PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 159/00, 15/18	A1	(11) International Publication Number: **WO 99/12118** (43) International Publication Date: 11 March 1999 (11.03.99)

(21) International Application Number: PCT/AU98/00715

(22) International Filing Date: 3 September 1998 (03.09.98)

(30) Priority Data:

PO 8921	3 September 1997 (03.09.97)	AU
PP 1192	31 December 1997 (31.12.97)	AU

(71) Applicants *(for all designated States except US)*: COMMONWEALTH SCIENTIFIC AND INDUSTRIAL RESEARCH ORGANISATION [AU/AU]; Limestone Avenue, Campbell, ACT 2061 (AU). MONASH UNIVERSITY [AU/AU]; Wellington Road, Clayton, VIC 3168 (AU).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: WINKLER, David, Alan [AU/AU]; Karanda, Cedar Grove, Belgrave, VIC 3160 (AU). BURDEN, Frank, Robert [AU/AU]; 23 Harrow Street, Blackburn South, VIC 3130 (AU).

(74) Agent: WATERMARK PATENT & TRADEMARK ATTORNEYS; 2nd floor, 290 Burwood Road, Hawthorn, VIC 3122 (AU).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, HR, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: COMPOUND SCREENING SYSTEM

(57) Abstract

The present application relates to a number of aspects of a compound and/or molecular screening system, including: 1) finding relationships between molecular structure and useful properties of molecules, more particularly using a virtual or mathematical analogue or model of a biological receptor or active site (a "virtual receptor") or other biological activity, such as toxicity; 2) creating a virtual receptor by use of Minimum Message Length or Maximum Entropy method (MEM) principles such as by applying a Bayesian regularised artificial neural network (BRANN); 3) using a virtual receptor to screen a database, the database may be real or virtual, may apply to existing or hypothetical molecules or compounds; 4) use of virtual receptors as fitness functions; 5) a method of mutating structures by modifying a SMILES string representation; 6) an improved molecular multipole moment representation; 7) an improved molecular eigenvalue index as a representation. Other aspects are also disclosed.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

COMPOUND SCREENING SYSTEM

FIELD OF INVENTION

The present application relates to a number of aspects of a compound and/or molecular screening system.

A first aspect of invention relates to the virtual screening of molecular representations, and in particular the invention is directed to the ability to evaluate the theoretical activity of molecules in various fields, such as, but not limited to, chemistry, agriculture (e.g. crop protection chemicals, growth modifiers), pharmacology (e.g. human and veterinary pharmaceuticals, toxicological profiles, diagnostic reagents) and the physical, physicochemical, and in particular biological activity of chemical compounds in general.

A further aspect of invention relates to refining the screening process in order to accentuate evaluation of likely active structures.

Still a further aspect relates to a method of mutating structures for evaluation by the screening system.

Still a further aspect relates to a fitness function which is used to assist in the evaluation of likely active structures.

Other aspects are also disclosed.

BACKGROUND OF THE INVENTION

The determination of the biological activity of chemical compounds is a continuing endeavour of research institutions and chemical companies, particularly due to its implications in the development of new drugs and other therapeutic remedies to treat or cure specific diseases. Biological activity of a compound is generally accepted as being the consequence of the fit of chemical compound into a receptor site involved in the particular biological process in a manner that the process is altered in some desirable way, e.g. either accentuated or inhibited.

Drug development generally starts with the discovery of a lead compound, which is a substance which exhibits a useful biological activity. Lead compounds are often obtained from natural sources or by synthesis of new chemical structures.

Traditionally the possible utility of compounds was determined by

screening the compounds to determine the degree of activity in a chemical or biological system of interest using in vitro or in vivo tests. This, however, proves to be very time consuming and expensive. The search for new active compounds has been speeded by the adoption of mass screening techniques utilising robots and other forms of automation and miniaturisation to perform and evaluate the tests; the throughput of such a system is many orders of magnitude greater than traditional screening. Synthesis of new potentially active compounds for screening has also been accelerated greatly by the adoption of the methods of combinatorial chemistry where many related compounds are made simultaneously. The mixed products of a combinatorial synthesis may be screened as a group or may be fully or partially separated into individual compounds before screening.

As powerful as these new methods of combinatorial chemistry and mass screening are, they are still only capable of accessing a very small region of the "universe" of possible chemical structures. For example, consideration of the numbers of possible branched chain isomers of alkanes shows that there are over 4 billion C_{30} isomers alone. Clearly more complex compounds with heteroatoms, rings and unsaturation are capable of being assembled in an almost infinite variety of ways. It is therefore not feasible, even using these new methods, to synthesise and screen more than a tiny fraction of the possible compounds for a single activity, let alone the many activities of practical interest and value.

Currently there are techniques available to generate mathematical representations of chemical compound structures. However, determining which of the many possible structures are most likely to exhibit useful biological activity of a particular kind, is problematic, i.e. which chemical structures out of the billions possible to synthesise and test in the search for active compounds.

A technique that has been developed to answer the problem of the relationship between chemical structure and activity is the Structure Activity Relationship (SAR) method which indicates which features of a molecule are responsible for a biological activity and which features do not play a role in its activity. In one form, SAR studies are usually based on structure/activity data

sets where there is a central structure common to all compounds but the substituents attached to this central structure are varied. If the compound's activity is unaffected by the variation, it is concluded that the substituent which was varied does not play an important role in the compound's mode of action. Conversely, if the compound's activity is affected by the variation, it is concluded that the modified substituent is involved in the compound's mode of action, whether that modification was favourable or not. This information may then be used as a guide to what types of analogues should be produced in the search for the most active compound by predicting the set of substituents which give the most active compound.

Further guidance in predicting the properties of a proposed lead compound is provided by the Quantitative Structure-Activity Relationships method.One way of determining the theoretically most highly active compounds is to use one of the various regression techniques to map molecular structure to activity, where the physicochemical properties are used to represent structure. This QSAR mapping allows determination of the values of the optimum physicochemical properties of the data set, and thus the structure of the most active compounds, may be determined. To determine the coefficients of the QSAR mapping an analytical technique is used, such as multiple linear regression (MLR). This type of QSAR modelling has its limitations due to higher order terms and cross product terms being neglected. To counter these inaccuracies, one improvement that has been made was to introduce "indicator variables" to the mathematical representation. Although such QSAR methods are capable of producing reasonable understanding of structure/activity relationships and can help discover more active compounds they have tended to be applied to small data sets concerned with a specific biological activity.

In the patent literature, PCT/CA96/00166, PCT/IB94/00257 and US 5,699,268 disclose inventions related to drug-receptor interactions. However in the case of PCT/CA96/00166, the embodiment of these simulated receptors is in a three dimensional, molecular level form. Therefore certain properties of the molecule as a whole are difficult, if at all possible, to ascertain. PCT/IB94/00257 discloses a method of calculating the free energy of binding of molecules to

receptors whose three dimensional structures have been determined by other means. US 5,699,268 discloses methods of generating computer simulated receptors using genetic evolution.

US 5,434,796 also discloses a computer simulated system for genetically evolving a population of molecules towards higher biological activity. The disclosure mainly relates to the way in which the generation of molecules for screening evolves. Basically, the disclosure revolves around the use of SMILES (Simplified Molecular Input LineEntry System) strings, which is described in "SMILES, a chemical language and information system. I. Introduction to methodology and encoding rules", D.Weininger, J. Chem. Inf. Comput. Sci., 28, 31 (1988). SMILES strings are lexical forms of molecular objects which are randomly mutated. However, the mutation rules are somewhat limited in that many types of chemically-important molecular modification are not readily accessible.

In the scientific literature, there are a number of documents, some of which disclose similar systems to those described above in the patent literature. In particular, Burden, F.R. *Quant.Struct.-Act. Relat.* (1996) 15, 7-11; Randic, M., *J.Amer.Chem.Soc.* (1975) 97,6609; Randic, M. and Trinajstic, N., *J.Molec.Struct.* (1993) 300,551-571; Kier, L.B. and Hall, L.H., *Molecular Connectivity in Structure-Activity Analysis*, J.Wiley and Sons, New York, 1986; Zhang, W., Koehler, K.F., Harris, B., Skolnick, P., Cook, J.M., *J. Med. Chem.* (1994) 37, 745-757; Harrison, P.W., *Eur. J. Med. Chem.* (1996) 31, 651-662; Davies,L.P., Barlin, G.B., Ireland, S.J., Ngu, M.M.L., *Biochem.Pharmacol.* (1992) 44, 1555-1561; Barlin, G.B., Davies, L.P.,Davis, R.A., Harrison, P.W., *Aust. J. Chem.* (1994) 47, 2001-2012; Fryer,R.I., Zhang, P., Rios, R., Gu, Z-Q, Basile, A.S., Skolnick, P., *J. Med. Chem.* (1993) 36, 1669-1673; Wang, C-G., Langer, T., Kamath, P.G., Gu, Z-Q.,Skolnick, P., Fryer, R.I., *J. Med. Chem.* (1995) 38, 950-957; Hollinshead, S.P., Trudell, M.L., Skolnick, P., Cook,J.M., *J. Med. Chem.* (1990) 33, 1062-1069; Allen, M.S., Hagen, T.J., Trudell, M.L., Codding, P.W., Skolnick, P., Cook, J.M., *J. Med.Chem.* (1988) 31, 1854-1861; and Yokoyama, N., Ritter, B., Neubert., A.D., *J. Med. Chem.* (1982) 25, 337-339.

The genetically evolved lead generation system draws on work done by

several groups which was aimed at generation of the large novel chemical databases referred to above (virtual combinatorial libraries). For example, Nilikantan, R, Bauman, N., Venkataraghavan, R.A. J.Chem. Inf. Comput. Sci. (1991) 31, 527-30 developed a method of random structure generation based on the random fusion of 2D chemical fragments. More recently, Clark, D.E., Firth, M.A., Murray, C.W. J. Chem. Inf. Comput. Sci. (1996) 36, 137-145 used graph theoretical techniques for vertex degree set generation and constructive enumeration of molecular graphs to generate 3D databases for drug design. Martin, Y.C. and van Drie, J.H. in *Chemical Structures 2. The International Language of Chemistry*, Warr, W.A.; Ed., Springer Verlag; Berling; (1993) pp 315-326 approached the problem by automatic modification of a SMILES string (Weininger D.J.Chem. Inf. Comput. Sci (1988) 28, 31-38) in order to meet specified geometrical constraints of a pharmacophore model. Modification of SMILES strings also features in the approaches of Mason J.S. in *Molecular Similarity in Drug Design*; Dean, P.M., Ed.; Blackie, Academic & Professional; London (1995); pp138-162 and Ho C.M.W. and Marshall, G.R. J. Comput.-Aided Mol. Des. (1995)9, 65-86 which generate hypothetical molecular structures to fit constraints, or generate large databases. SMILES strings are converted into structures using programs such as CONCORD (Rusinko, A., Sheridan, R.P., Nilakanta, R., Haraki, K.S., Bauman, N., Venkataraghavan, R. J.Chem. Inf. Comput. Sci. (1989) 29, 251-255). Also Glen, R.C. and Payne, A.W.R., J. Comput.-Aided Mol Des. (1995) 9, 181-202 use a genetic algorithm to again generate novel, structurally-diverse 3D databases, although no detailed illustrations of its use are given.

Despite all that is disclosed in the literature noted above, there is still a need for a method of accessing very large numbers of structures, which also enables assessment of these chemical compounds and prediction of their likely affinity for given receptor sites with relatively increased throughput, simplicity and flexibility over the prior art.

SUMMARY OF THE INVENTION

The present application relates to a number of aspects, including:

1. finding relationships between molecular structure and useful properties of

molecules, more particularly using a virtual or mathematical analogue or model of a biological receptor or active site (a "virtual receptor") or other biological activity, such as toxicity ;

2. creating a virtual receptor by use of Minimum Message Length or Maximum Entropy Method (MEM) principles, such as by applying a Bayesian regularised artificial neural network (BRANN);
3. Using a virtual receptor to screen a database, the database may be real or virtual, may apply to existing or hypothetical molecules or compounds;
4. Use of virtual receptors as fitness functions;
5. A method of mutating structures by modifying a SMILES string representation;
6. An improved molecular multipole moment representation;
7. An improved molecular eigenvalue index as a representation.

These aspects are described in more detail in the following sections.

It is to be noted that in this specification, the term "compound or molecule" is taken to include parts of compounds or molecules.

1. Virtual analogue of a receptor ("virtual receptor")

This aspect provides a method of creating a virtual receptor capable of being used to scan a range of compounds and providing a measure indicative of whether the compounds are likely to exhibit a particular characteristic, including the steps of:

- compiling a data set of compounds which exhibit the known characteristic;
- forming a conceptual structure/activity model with a given architecture;
- converting the data set into a representation readable by the conceptual model;
- training the conceptual model on at least a portion of the converted data set in order to improve the architecture of the conceptual model.

Preferably, the data input to the virtual receptor is a molecular representation of the compounds which include the entire molecule and

embody relevant properties such as steric, electronic and lipophilic properties. This makes it possible for embodiments of the invention such as those employing neural networks to model the activity of structurally diverse data sets. It also avoids the difficulty of finding the values of physicochemical properties which give the most highly active compound. A preferred output of a virtual receptor that may be determined is the binding affinity of the compounds or other biological activity.

The details of the use of neural networks to create a virtual receptor, and the types of molecular representation which might be used in the virtual receptor training and compound screening is discussed with reference to the detailed description following.

Artificial Neural Networks

A further aspect is based on the use of a mathematical concept called an artificial neural net to derive a virtual receptor. Artificial neural networks (ANNs) are mathematical models, and thus it has been found that they can be used in respect of scanning compounds and training virtual receptors.

Preferably, an evolutionary neural network may be used.

Molecular Representations

The virtual receptor may be rendered in a number of forms. Preferably the rendering is in a mathematical form. One form may be by the atomistic approach, which classifies each atom according to its element and the number of connections.

Another form of a molecular representation is achieved by dividing the molecules being input into the ANN into functional groups capable of interacting with receptors, namely: CO2-, PO42-, N+, N, OH, C=O, O/S ethers, halogens, Csp3, Csp2 and an entropic term related to the number of freely rotatable bonds in the molecule (DOF).

Alternatively, the compounds may be represented in terms of simple molecular structural parameters, such as constituent atoms or functional groups.

An advantage that stems from the inventive method using an atomistic representation is that it allows compounds to be screened with no more knowledge than is provided by counting molecular fragments.

Many other molecular representations however are possible, such as depicting the molecules based on their optimal physicochemical properties (see example 2 below). In addition, topological indices, Burden's chemically intuitive molecular index (CIMI), and/or molecular hologram representation of Tripos Assoc. may be used as compound descriptors. Additional novel representations which form additional aspects of the invention are exemplified in the sections following.

It is to be noted that one inventive concept involves the creation of a virtual receptor by training the receptor using compounds with known properties. Once a virtual receptor has been created based on a particular molecular or mathematical representation of the compounds, all future compounds that are used as input to that receptor must also be represented in the particular molecular or mathematical representation used in the training of the receptor.

2. Virtual receptor generation using a Bayesian regularised artificial neural network (BRANN)

This aspect provides a method of generating a virtual receptor by use of models which exhibit stability or compensate for noise. One such model is a Bayesian regularised artificial neural network (BRANN). Another model is Maximum Entropy Method (MEM).

Using Bayesian regularisation (MacKay, 1992) removes the need to supply a validation set since it minimises a linear combination of squared errors and weights. It also modifies the linear combination so that at the end of training the resulting network has good generalisation qualities. It has also been suggested that there is no need for a testing set since the application of the Bayesian statistics provides a network that has maximum generalisation. Concerns about overfitting and overtraining are also removed by this method so that the production of a definitive and reproducible model is attained.

3. Using a virtual receptor to screen a database

The usefulness of this invention is particularly apparent when recent developments in the generation of chemical databases, discussed in the background to the Invention, are considered. These databases represent virtual combinatorial libraries. They may be completely random or biased

around core structures/chemistry. These large databases of hypothetical and/or real compounds may be converted into a molecular or mathematical representations of the compounds and used input to a virtual receptor.

The present aspect may also be used to screen databases or chemical libraries of real, synthesised compounds derived using the concept of combinatorial chemistry.

Another aspect of the invention, which may be referred to as a Virtual Screening Process, is predicated on the discovery that by creating a "virtual receptor" first, and then using this virtual receptor to screen compound libraries,it is possible to test, in a "virtual" environment, the compatibility of each compound being screened to the virtual receptor.

Thus, this aspect provides a method of screening a range of compounds, including:

(a) creating a virtual or mathematical analogue of a biological receptor or active site (a "virtual receptor");

(b) scanning a range of possible compounds as their mathematical representations with the virtual receptor so as to provide a measure of the likelihood of each compound exhibiting a particular characteristic.

A preferred measure that may be determined is the binding affinity of the compounds or other biological activity.

If a given compound contains certain structural features (i.e. conforms to a pharmacophore) there is a high likelihood of the compound having a particular biological activity. Due to the screening being done in a "virtual" environment, the need to synthesise a large number of compounds is avoided. The number of compounds synthesised is reduced to those predicted as being suitable in the "virtual" environment, and which also have a higher likelihood of being verified in the real world.

In another preferred form, the virtual receptor is continually modified, in order to improve its prediction abilities, based on compounds located in database scans that have proved to in fact exhibit the characteristics sought.

As discussed previously a preferred form of this "virtual environment" is a neural network in a computer environment. Hardware implementations of neural nets

are also possible (and may be preferable once a virtual receptor of a given type is defined and large databases are to be screened).

4. Genetic evolution of structures using virtual receptors as fitness functions

Whilst the generation of virtual combinatorial libraries is a rapidly developing field of research, it is acknowledged that literal simulation of the combinatorial chemistry/high throughput screening process in silico will require lengthy searches of the vast combinatorial space encompassed by the libraries. Although this will allow searching of the universe of chemical compounds which are theoretically possible orders of magnitude more quickly than can real world combinatorial synthesis, it can still only explore a minute fraction of chemical space possible. An alternative, complementary approach is the subject of this further inventive disclosure. It has advantages in that it may be much more efficient at exploring chemical space than virtual screening methods.

Additional aspects of the invention include the use of virtual receptors as fitness functions, and the discovery of efficient methods of mutating chemical structures to span as much of combinatorial space as possible. The aspect of the invention involving mutation strategies is discussed in the next section.

We disclose a process where the virtual receptor of specified type is used as a 'fitness function' in a genetic algorithm which evolves a population of chemical structures. This population can be biased in order to capitalise on a particular type of chemistry, or may be completely random. It is envisaged that diverse types of considerations would be incorporated into the fitness function. For example, evolved structures would need to be screened to ensure they represent chemically sensible compounds. They may also be screened for ease of synthesis, toxicity etc using other algorithms. The system allows one or more characteristics to be selected sequentially or in combination during the optimisation.

5. Mutating structures by modifying a SMILES string

Mutation Strategies

A mutation operator determines that, with some low probability, a portion of the new individuals will have some of their bits flipped.

In a crossover operation, two individuals are chosen from the population using a selection operator.

This aspect provides using mutation and cross-over strategies as applied to SMILES strings, in order to modify the behaviour of the SMILES string as applied to a compound screening system.

6. Improved molecular multipole moment representation

Ultimately, of a virtual receptor is dependent on the quality of the molecular representation used to develop it. The quality of the virtual receptor is also dependent on the quality of the training data and possibly on the architecture of the neural net. In this regard, it is preferred that the numerical representation of the compound being analysed adequately represents the steric, electronic and lipophilic properties of the whole molecule.

A preferred representation is the molecular multipole moment (MMM) representation, which is achieved by generating the zero-, first- and second-order molecular multi-pole moments(MMM) with respect to atomic mass and/or atomic charge.

We disclose an improved method of generating molecular multipole moments as now described. In order to generate electrostatic, lipophilic or any other set of multipole moments, it is necessary to define a consistent set of axes. These axes should be invariant to a change in the elemental nature of an atom at any particular site, e.g. replacement of H by F. The axes should only vary their position or orientation when there is a structural change in the molecule, eg change in molecular conformation, an increase or decrease in the number of atoms making up the molecule, or some other structural change.

7. Improved molecular eigenvalue index representation

The further aspect of the invention is an additional type of molecular representation. It involves the generation of useful molecular descriptors from eigenvalues of adjacency, or modified adjacency matrices in which the diagonal elements are values relating to steric, electrostatic or lipophilic properties of the constituents atoms of the compounds. In a preferred embodiment it is envisaged that eigenvalues of three matrices (one each of steric, electrostatic, and lipophilic-related properties) would be generated. The steric diagonal

elements of the adjacency, or modified adjacency matrices could be the Vander Waals radii of the atoms; the electrostatic diagonal matrix elements could be the atom charges derived from empirical or molecular orbital calculations and; the lipophilic diagonal matrix elements could be the atomistic lipophilicities referred to in the section above on molecular multipole moments.

DETAILED DESCRIPTION OF THE INVENTION

Preferred embodiment(s) of the inventions disclosed above will now be described, with reference to the accompanying drawings,where:

Figure 1 shows a set of data used in an example,

Figure 2 illustrates an example size of training, validation sets and number of networks generated,

Figure 3A illustrates an example measure of the predictive ability of a network,

Figure 3B illustrates the B5 representation,

Figure 3C illustrates a summary of the A1 representation,

Figure 3D illustrate results obtained for the MJ representation,

Figure 4A illustrates a sample output from a 23:2:1 neural network using the B2 representation as input,

Figure 4B illustrates a sample output from an 11:4:1 neural network using the B3 representation as input,

Figure 4C illustrates an sample output of 11:4:1 neural network using A1 representation as input,

Figure 5 shows an optimal architecture,

Figure 6 shows results for example 3,

Figure 7 shows a sample output from a 21:8:5:3:1 network,

Figure 8 shows a comparison of neural network and MLR,

Figure 9 shows the results of example 4,

Figure 10 shows an example flowchart of a genetically-evolved lead generation system as disclosed in accordance with the further disclosed 'fitness function' invention,

Figure 11 illustrates a summary of the genetic algorithm,

Figure 12 illustrates an example mutation operator, and

Figure 13 illustrates an example cross-over operator,

Figure 14 illustrates an overall concept flowchart for virtual receptor generation.

Figure 15 illustrates a virtual screening flowchart showing use of virtual receptor to predict properties of library members, library can be real or virtual.

Figure 16 illustrates a genetically evolved chemical library overview flowchart.

Figure 17 illustrates a genetically-evolved chemical library detailed flowchart showing role of fitness functions and specific examples of smiles mutation.

Figure 18 illustrates a flowchart of improved multipole moment molecular representation generation.

Figure 19 illustrates a flowchart for generation of improved eigenvalue indices as molecular representations.

Figures 20 illustrates Muscarinic virtual receptor training, observed versus calculated scaled log (activity) for training set (examples).

1. Virtual analogue of a receptor ("virtual receptor")

As summarised above, there is provided a method of screening a range of compounds which includes

(a) creating a virtual or mathematical analogue of a biological receptor or active site (a "virtual receptor") and

(b) scanning a range of possible compounds as their mathematical representations with the virtual receptor so as to provide a measure of the likelihood of each compound exhibiting a particular characteristic.

A preferred measure that may be determined is the binding affinity of the compounds or other biological activity.

This method of creating a virtual receptor may also be used to scan a range of compounds and provide a measure indicative of whether the compounds are likely to exhibit a particular characteristic in which it includes the steps of:

- compiling a data set of compounds which exhibit the known characteristic;

- forming a conceptual structure/activity model with a given architecture;
- converting the data set into a representation readable by the conceptual model;
- training the conceptual model on at least a portion of the converted data set in order to improve the architecture of the conceptual model.

Preferably, the data input to the virtual receptor is a molecular representation of the compounds which consider the entire molecule and embody relevant properties such as steric, electronic and lipophilic properties. This makes it possible for embodiments of the invention such as those employing neural networks to model the activity of structurally diverse data sets. It also avoids the difficulty of finding the values of physicochemical properties which give the most highly active compound. A preferred output of a virtual receptor that may be determined is the binding affinity of the compounds or other biological activity.

The details of the use of neural networks to create a virtual receptor, and the types of molecular representation which might be used in the virtual receptor training and compound screening are discussed below with reference to Figure 14.

Artificial Neural Networks

Virtual receptors can be generated by a number of different methods, many of which rely essentially on regression in one form or another. A particularly useful way of deriving a virtual receptor is to use a mathematical concept called an artificial neural net. Artificial neural networks (ANNs) provide an improved platform from which to predict the behaviour of molecules. Several advantages in using neural networks are that they are fast, they do not rely on subjective judgements as to the form of the functional relationships between structure and activity to be provided, and they process numerous parameters simultaneously. In addition, they are robust and capable of producing reasonable results even when the data is noisy.The prime advantage of using neural networks over other known methods, however, lies in their ability to internally process complex non-linear relationships.

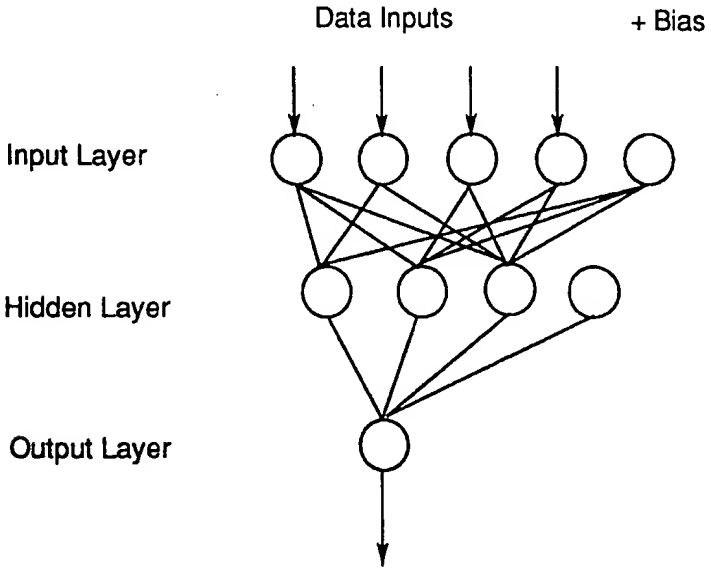
ANNs are mathematical models, based loosely on the way biological neural networks process information. ANNs consist of layers of artificial neurones (or neurodes). Each neurode has numerous inputs (x1,x2,...) each of which is modified by a weight (w1,w2,...). These inputs are summed on entry to the neurode. This net input is then modified by an internal transfer function.

Examples of transfer functions that may be used are linear step

functions,which pass a signal if the net input exceeds a certain threshold, or a function that produces a continuous, differentiable non-linear signal, such as a sigmoidal or hyperbolic tangent functions (otherwise known as "squashing functions").

The output of the internal transfer function forms the output of the neurode, which is either passed on as the input for other neurodes or as an output carrying a result.

The structure or architecture of ANNs can take many forms, such as single layer, multi-layer, feed forward and lateral connectivity. In these various architectures, the layers of neurodes may be fully or partially connected. A full connection is where the output of a neurode is passed onto each neurode in the next layer, whereas in a partial connection the output is transferred only to selected neurodes. An example of a three layered 4:3:1 ANN architecture is shown below:



The output of an ANN depends upon numerous factors, namely the nature of the neurodes' transfer functions, the architecture of the network and the weights connecting the neurodes. Of these factors, the weights connecting the neurodes are most easily altered.

The ANN as a whole is trained so that it is capable of recognising the important characteristics in molecules that may mean that they exhibit a

biologically useful activity. That is, the representations of molecules, with known properties, are repeatedly input to the ANN. The ANN is then modified by adjusting the weights connecting the neurodes until the error between its outputs and the correct outputs is minimised. The method used to adjust the weights in the process of training the ANN is called "the learning rule" and may be supervised or unsupervised. Back propagation is an example of a supervised learning rule.

Back propagation is a gradient descent algorithm. The network error may be considered a function of the network weights. Back propagation minimises the average squared error between the network output and the "correct answer" by moving down the gradient of this error function. The network weights are altered according to the Delta Rule (also known as the Least Mean Squared Rule). In training a back propagation ANN, the output is compared with the desired result, and a proportion of this error determined is then propagated back through the network, with the network weights modified accordingly.

Although an ANN's ability to model the data upon which it is being trained increases with the length of the training procedure, this is not true regarding a network predictive capabilities. An ANN's ability to produce the correct output for input patterns outside its training data set will improve initially, as the network learns rules relating input patterns to output, however, after a certain number of training cycles, the network will begin to "memorise" the training data - that is whilst its ability to give accurate outputs of the training data set will continue to improve,the network's ability to generalise will diminish. To minimise this problem, it is therefore necessary to include a validation step in the testing routine in order to ensure that a suitable virtual receptor is created. The set of weights which give the lowest validation error are said to be the network's "best weights" as the network's predictive abilities are greatest at this point.

When using ANNs it is necessary to find the ANN architecture which best solves the problem at hand. The number of neurodes in the input layer and the output layer will be determined by the number of input parameters and the number of outputs respectively. However, ascertaining the optimal number of hidden layers (the layers between the input and the output layers) and the

number of neurodes in those layers is not as simple.

It is to be noted that recent research may eliminate the trial-and-error aspect of network optimisation, as a new system called the evolutionary neural network uses a genetic algorithm to optimise not only the network topology, but also the set of input parameters used. This automatically optimises the network architecture and chooses the best set of input parameters. We discuss in the next section an aspect of the invention involving a novel method for removing this problem of optimisation of neural network architecture.

Molecular Representations

There are a number of established methods by which molecules may be represented mathematically for use in structure-activity mapping methods. One embodiment of a molecular or mathematical representation of compounds is the atomistic approach, which classifies each atom according to its element and the number of connections. For instance a carbon atom with four connections is denoted C4, those with three connections C3 etc and the number of atoms of each type totalled. Although simple, this representation is adequate to encode not only physicochemical parameters such as lipophilicity and molar refractivity, but also biological activity. Table 1 provides some coding examples:

Table 1. An example of atomistic encoding of molecules.

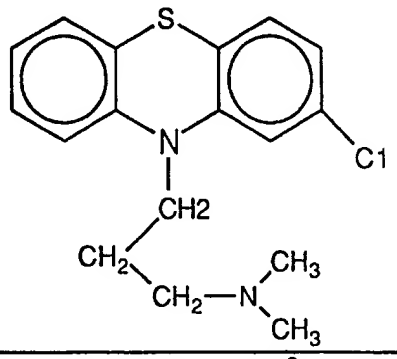
TABLE 1

Molecule	H	C4	C3	C2	O2	O1
Ethanol, C2H5OH	6	2	0	0	1	0
Diethylether C2H5OC2H5	10	4	0	0	1	0
Acetone, CH3COCH3	6	2	1	0	0	1

Another embodiment of a molecular representation is achieved by dividing the molecules being input into the ANN into functional groups capable of interacting with receptors, namely: CO2-, PO42-, N+, N, OH,C=O, O/S ethers, halogens, Csp3,Csp2 and an entropic term related to the number of freely rotatable bonds inthe molecule (DOF). Research involving molecules of diverse structures has shown that these functional groups play an important role in the ability of molecules to interact with receptors. In particular, it has been shown that this functional group representation is able to encode

physicochemical parameters, such a lipophilicity and molar refractivity, as well as biological activity. The fact that steric and lipophilic factors are often important in drug receptor interactions provides a partial explanation as to why such a simple representation is capable of being used as input to the virtual receptor of the invention. An example of how this numeric achieved is shown below in Table2

TABLE 2 Functional group representation of chlorpramazine.



DOF[a]	Csp^2	Csp^3	N^+	N	CO_2^-	PO_4^{2-}	OH	C=O	O/S	Hal
4	12	5	0	2	0	0	0	0	1	1

[a] Degrees of Freedom

Alternatively, the compounds may be represented in terms of simple molecular structural parameters, such as constituent atoms or functional groups.

An advantage that stems from the inventive method using an atomistic representation is that it allows compounds to be screened with no more knowledge than is provided by counting molecular fragments.

Many other molecular representations however are possible, such as depicting the molecules based on their optimal physicochemical properties (see example 2 below). In addition, topological indices, Burden's chemically intuitive molecular index (CIMI), and/or molecular hologram representation of Tripos Assoc. may be used as compound descriptors. Additional novel representations which form further aspects of the invention are exemplified in the sections following.

The output generated by the virtual receptor upon screening a range of compounds would indicate which compounds have the highest likelihood of forming the basis of new lead compounds. The most novel of these could also be used to synthesise biased combinatorial libraries of organic compounds for screening in pharmacological receptor assays. The use of a neural network to map structure to activity results in superior models to the use of linear methods such as MLR or PLS. This reflects the presence of non-linear relationships between structural parameters and activity, and interactions between the descriptors. The ability of neural networks to account for these relationships is an advantage in virtual receptor generation.

It is to be noted that the inventive concept involves the creation of a virtual receptor by training the receptor using compounds with known properties. Once a virtual receptor has been created based on a particular molecular or mathematical representation of the compounds, all future compounds that are used as input to that receptor must also be represented in the particular molecular or mathematical representation used in the training of the receptor.

2. Virtual receptor generation using a Bayesian regularised artificial neural network (BRANN) or Maximum Entropy Method (MEM)

Regression is an "ill-posed" problem in statistics, which sometimes results in structure-activity models exhibiting instability when trained with noisy data. Regression methods, including back propagation neural nets, also face additional problems. Principal amongst these are overtraining, overfitting, and selection of the best QSAR model from a number obtained in the validation process. Overtraining results from running the neural network training for too long and results in a loss of ability of the trained net to generalise. Overtraining can be avoided by used of a validation set. Cross-validation, which provides a good test for the predictive capabilities of a network, also provides assistance in determining the optimal neural net architecture. Cross-validation involves running a data set through a network numerous times until all data points have been in both the training and the validation sets.

Overfitting results from the use of too many adjustable parameters to fit the training data and is avoided by use of test sets of data, not used in the training and validation. However, the validation of neural networks (and most regression methods which might be used to generate virtual receptors) scales as O(N2P2), where N is the number of data points and P is the number of input parameters, are exacerbated by very large data sets. For virtual screening, where the training data set should be comprehensive and diverse, this can result in a very large validation time and effort, and is potentially a serious problem.

An aspect of the invention, in which the problems of model instability, lengthy cross validation processes, and the ad hoc methods of network architecture optimisation may be simultaneously solved, involves the use of a Bayesian regularised neural network. This belongs to the general class of of Minimum Message Length (MML) methods and similar results may be obtained by alternative methods such as the Maximum Entropy Method (MEM). However, the Bayesian regularised artificial neural network (BRANN) may be better suited to virtual receptor calculations than other regression methods. Neural network training can be regularised, a mathematical process which converts the regression into a well-behaved "well-posed" problem and overcomes model instability. Bayes theorem provides the correct language for describing the inference of a message communicated over a noisy channel. In structure-activity models the 'noise' corresponds to experimental error, poor choice of molecular representations etc. The SAR 'message' corresponds to a useful, valid structure-activity model (or virtual receptor). Where orthodox statistics provide several models with several different criteria for deciding which model is best, Bayesian statistics only offers one answer to a well-posed problem.

The advantage of BRANN is that the models are robust and the potentially lengthy validation process discussed above, is unnecessary. Using Bayesian regularisation (MacKay, 1992) removes the need to supply a validation set since it minimises a linear combination of squared errors and weights. It also modifies the linear combination so that at the end of training

the resulting network has good generalisation qualities. It has also been suggested that there is no need for a testing set since the application of the Bayesian statistics provides a network that has maximum generalisation. Concerns about overfitting and overtraining are also removed by this method so that the production of a definitive and reproducible model is attained.

3. Using a virtual receptor to screen a database

The usefulness of this invention is particularly apparent when recent developments in the generation of chemical databases, discussed in the background to the Invention, are considered. These databases represent virtual combinatorial libraries. They may be completely random or biased around core structures/chemistry. These large databases of hypothetical and/or real compounds may be converted into molecular or mathematical representations of the compounds and used as input to a virtual receptor. The present invention may also be used to screen databases or chemical libraries of real, synthesised compounds derived using the concept of combinatorial chemistry.

Another aspect of the invention, which may be referred to as a Virtual Screening Process, and one embodiment of which is illustrated in Figure 15, is predicated on the discovery that by creating a "virtual receptor" first, and then using this virtual receptor to screen compound libraries, it is possible to test, in a "virtual" environment, the compatibility of each compound being screened to the virtual receptor. If a given compound contains certain structural features (i.e. conforms to a pharmacophore) there is a high likelihood of the compound having a particular biological activity. Due to the screening being done in a "virtual" environment, the need to synthesise a large number of compounds is avoided. The number of compounds synthesised is reduced to those predicted as being suitable in the "virtual" environment, and which also have a higher likelihood of being verified in the real world.

In another preferred form, the virtual receptor is continually modified, in order to improve its prediction abilities, based on compounds located in database scans that have proved to in fact exhibit the characteristics sought.

As discussed previously a preferred form of this "virtual environment" is

a neural network in a computer environment. Hardware implementations of neural nets are also possible (and may be preferable once a virtual receptor of a given type is defined and large databases are to be screened).

4. Genetic evolution of structures using virtual receptors as fitness functions

Whilst the generation of virtual combinatorial libraries is a rapidly developing field of research, it is acknowledged that literal simulation of the combinatorial chemistry/high throughput screening process in silico will require lengthy searches of the vast combinatorial space encompassed by the libraries. Although this will allow searching of the universe of chemical compounds which are theoretically possible orders of magnitude more quickly than can real world combinatorial synthesis, it can still only explore a minute fraction of chemical space possible. An alternative, complementary approach is the subject of this further inventive disclosure. It has advantages in that they may be much more efficient at exploring chemical space than virtual screening methods.

Additional aspects of the invention include the use of virtual receptors as fitness functions, and the discovery of efficient methods of mutating chemical structures to span as much of combinatorial space as possible. The aspect of the invention involving mutation strategies is discussed in the next section.

In our further invention these algorithms are used to generate new 'child' populations of structures in a genetic algorithm. Each structure is mutated by means of single point mutations, insertions, deletions and crossovers, to generate another population of structures for testing against the fitness function represented by a virtual receptor and possibly others such as ease of synthesis, toxicity etc. Examples of library evolution are shown in Figures 10, 16 and 17. The aspect considered unique to the approach is that the mutated structures together with a suitably defined fitness function and evolutionary process, such a genetic algorithm or other types of genetic programs, can be used to explore very large areas of combinatorial space and generate lead structures likely to be active at the specified receptor. With

suitable mutation strategies, as can be defined by the task at hand, virtually any part of the combinatorial universe (the region of compound space theoretically allowed by the rules of chemical bonding) can be explored. This means that areas devoid of useful biological activity would be relatively rapidly vacated by the genetic algorithm, with more promising regions developing larger populations of lead structures. By reducing the level of similarity which will be tolerated, it may be possible to drive the population of chemical structures to high diversity whilst retaining useful lead activity.

Evolutionary Processes

Genetic algorithms are a well known form of genetic process and are automated heuristics that perform optimisation by emulating Darwinian "Survival of the Fittest". Further disclosure is made for ease of reference to a genetic algorithm, but the invention is not limited only to this. In this scheme, potential solutions to our problem are analogous to "individuals". Each individual is made up of a set of values, or "genes", that define the individual's characteristics. How well an individual solves the optimisation problem is the individual's "fitness". An individual's fitness is important because it determines the individual's likelihood for survival and mating.

In one implementation, of the many possible, and as illustrated in Figure 11, the algorithm starts with an initial population of these individuals. The fitness of each is evaluated to determine how well it solves the problem. Typically the characteristics of each individual in the initial population are generated randomly. Next, two individuals are selected from the population. This is done so that the individuals that are more fit are more likely to be selected. The two selected individuals can be considered to be "parents". From the parents, two new individuals ("children") are created that are recombinations of the genes from the parents. The process of creating the children is called "crossover" Some combination of the parents and children are then passed to the "next generation".

The selection and crossover steps are repeated until the number of individuals in the next generation is the same as that in the current generation. That is where mutation comes in. A small number of genes in the population

are replaced with randomly generated values. This has the function of injecting into the population potentially good gene values, or characteristics, that may not have occurred in the initial population, or that may have been lost through selection, crossover and mutation processes.

At this point, the algorithm may use a process called "elitism" in which some of the best organisms from the previous generation replace some of the worst individuals in the current generation. This has less of an analogy with real life, but assures that the best individuals do not disappear.

In the new generation, good genes - and good combinations of genes - are typically more likely to occur than in the previous generation. So, with each generation of the model, the solutions generally do a better and better job of solving the problem.

In practice, however, we can implement this genetic model of computation by having arrays of bits or characters to represent the chromosomes. Simple bit manipulation operations allow the implementation of crossover, mutation and other operations. The genetic algorithm (as a simulation of a genetic process) is not a random search for a solution to a problem (highly fit individual). The genetic algorithm uses stochastic processes, but the result is distinctly non-random (better than random).

Potential Selection Operators

A selection operator is usually used to select which member of an evolving population will be involved in crossover or other mutations. In human terms this may be analogous to selection processes which favour the most powerful male mating with the most desirable female. In this application to lead discovery selection operators choose which two or more molecules will be involved in crossover or other mutations. These operators may be:

- selecting the best and second best molecules for crossover; or
- randomly choosing two molecules out of the best n; or
- selecting the best and next best with a molecular similarity less than some cutoff value; or
- choosing the best members of two or more evolving populations of molecules etc.

Potential Fitness Functions

A selection operator is used to give preference to better individuals, allowing them to pass on their genes to the next generation. The goodness of each individual depends on its fitness, which may be determined by an objective function or by a subjective judgement. A 'global' fitness function may involve either a weighted average of some or all of component functions, or some of the fitness criteria may be applied sequentially. An example of the sequential application is for all members of the evolving populations(s) may have their fitness evaluated against the chemical valence fitness function (to eliminate nonsense compounds) then be evaluated for biological activity fitness via the virtual receptor. The most active molecules as determined by the virtual receptor fitness function may then be 'filtered' for toxicity or some other property.

We disclose a process where the virtual receptor of specified type is used as a 'fitness function' in a genetic algorithm which evolves a population of chemical structures. This population can be biased in order to capitalise on a particular type of chemistry, or may be completely random. It is envisaged that diverse types of considerations would be incorporated into the fitness function. For example, evolved structures would need to be screened to ensure they represent chemically sensible compounds. They may also be screened for ease of synthesis, toxicity etc using other algorithms. The system allows one or more characteristics to be selected sequentially or in combination during the optimisation.

In our invention fitness functions may be exemplified by some of the following types (not an exhaustive list):

- A valence function which determines whether the structure represented by the chromosome obeys the laws of chemical bonding and valence. For example, it would eliminate structures without four bonds to carbon, with isolated aromatic atoms next to aliphatic etc. This type of fitness function can be adapted from the algorithms used to parse SMILES strings.

- A stability function which eliminates chemically unstable or extremely difficult to synthesise structures such as peroxides, or large numbers of chiral centres. This could be derived from a lookup table of undesirable functional groups.
- A safety function which rates the structures represented by the chromosomes in terms of likely toxicity. For example, nitrogen mustards, alkylating agents etc would be eliminated. This could be derived from structure-activity models in a similar way to the Topkat commercial software.
- A biological activity function. This would be implemented via the virtual receptor concept as disclosed above. It is most likely implemented as a neural network model.
- A molecular diversity function. The evolutionary algorithms used in this invention have a stochastic element which ensures a degree of molecular diversity. However, another fitness function would be used which ensures that, for example, no individual in the population has a greater than 85% similarity to the others. This function may also screen out molecular redundancies.
- "ease of synthesis" fitness. There are empirical algorithms to evaluate this in some commercial modelling packages (e.g. Leapfrog from Tripos Associates). This is based on the number of chiral centres, rings etc.
- "biological selectivity" fitness. This is a measure of the molecule's ability to affect the desired receptor and not another (possibly closely related one). For example, leads for Alzheimer's disease therapy should affect the M1 muscarinic receptor but not the M2 to M5 receptors as peripheral side effects may result.
- "cost" fitness. This is a more important factor for agrochemicals than drugs. It is possible to estimate the cost of synthesis using empirical relationships.
- "metabolic liability" fitness. This is a measure of how quickly a molecule is inactivated by metabolism. Algorithms to estimate this are available in commercial systems eg MetabolExpert.

- "combinatorial synthesis" fitness. The fitness function may determine whether combinatorial methods may be adapted to be used in the synthesizing compounds for screening.
- "pharmacokinetic efficiency" fitness. This is a measure of how well the molecule is transported from its site of entry to the site of action. A simple example of this may be whether a CNS active drug can penetrate the blood-brain barrier.
- "chemical advantage" efficiency. i.e. does the class of molecules tap into any special chemical expertise resident in the research group which is implementing the search?

A further aspect of the invention is based on the concept of using evolutionary modification of compound structures whereby the calculated activity from the Virtual Screening Process is used as a measure of the 'fitness' of a chemical structure for performing a particular function. The better, or a predetermined group of, compounds can be selected based on the 'fitness' or arrange of 'fitness' as base structures for subsequent genetic modification. 'Fitness' may be considered as an assessment of a compound exhibiting survival of the fittest in a genetic algorithm. Optimisation provides a'fitness function'. The fitness function is used to evaluate the "fitness", or superiority of one member of a population over another by some definable criteria. In this application, the fitness function is the mathematical embodiment of the criteria used to define the "fitness" of a chemical compound over another. The criteria can be set according to the particular result required or outcome hoped for.

Variations and additions of the inventions disclosed are possible within the general inventive concept as will be apparent to those skilled in the art.

5. Mutating structures by modifying a SMILES string

Mutation Strategies

The mutation operator determines that, with some low probability, a portion of the new individuals will have some of their bits flipped. An example is shown in Figure 12.

Its purpose is to maintain diversity within the population and inhibit

premature convergence. There is relationship between the bit string and a molecular structure, which is usually 1:1 (except in some cases where optical or geometric isomers are not accounted for). It may be noted that molecular structures may not literally be represented by bit strings but the same operations and logic which apply to bit strings in the general discussion of genetic algorithms will also apply to other representations of molecules. It should be possible, for example, to use the SMILES string to represent a molecule, then alter this by symbol substitution, addition, fragment insertion or deletion etc to produce evolved structures via the genetic algorithm and the fitness function. Mutation alone induces a random walk through the search space. Mutation and selection (without crossover) create a parallel, noise-tolerant, hill-climbing algorithm.

In the crossover operation, two individuals are chosen from the population using a selection operator. A crossover site along the bit strings is randomly chosen. The values of the two strings are exchanged up to this point. An example is shown in Figure 13. If S1 = 000000 and S2 - 111111 and the crossover point is 2 then S1'= 110000 and S2' = 001111. The two new offspring created from the mating are put into the next generation of the population. By recombining portions of good individuals, this process is likely to create even better individuals.

The crossover operation happens in an environment where the selection of who gets to mate is a function of the fitness of the individual, i.e. how good the individual is at competing in its environment. Some genetic algorithms use a simple function of the fitness measure to select individuals (probabilistically) to undergo generic operations such as crossover or asexual reproduction (the propagation of genetic material unaltered). This is fitness-proportionate selection. Other implementations may use a model in which certain randomly selected individuals in a subgroup compete and the fittest is selected. This is called tournament selection and is the form of selection we see in nature when stags rut to vie for the privilege of mating with a herd of hinds. The two processes that are considered to most contribute to evolution are crossover and fitness based selection/reproduction. As it turns out, there

are mathematical proofs that indicate that the process of fitness proportionate reproduction is, in fact, near optimal in some senses.

The following are examples of potentially useful mutation operators for SMILES strings. The choice of which mutation operator is carried out on a given member of the chemical population can be decided randomly eg by use of a number wheel algorithm.

Insertion: Insertion mutations involve randomly selecting a character position in the string and inserting one or more chemically parsable text strings at that position. The choice of which string to insert could, for example, be chosen randomly from a large lookup table of SMILES strings. Some of the strings in the lookup table, or other selection process which derives the string to be substituted, could be contained in brackets. In this case the insertion results in a branching of the new string from the old. Strings inserted without these enclosing brackets would be incorporated into the original molecule without branching.

e.g. original string CCCCCC
mutated string CCCSCCC (chain insertion)
e.g. original string CCCCCC
mutated string CCC(CCOC)CCC (chain branch)

Deletion: As with insertion, a position in the SMILES representation of the molecule to be mutated is chosen at random. One or more atoms (number chosen randomly within a range) are removed from the string to generate a new string.

e.g. original string CCCCCC
mutated string CCCCC

Substitution: As with the previous operators, a position in the SMILES string is chosen randomly and one or more characters are swapped for different characters. When more than 2 characters are substituted, this would most likely be done with a lookup table so as to preserve chemical reasonableness.

e.g. original string CCCCCC
mutated string CCCSCC (simple substitution)

e.g. original string CCCCCC

mutated string CCC(CC(=O)NCC)CC (branch substitution)

In our application these algorithms are used to generate new child populations of structures in a genetic algorithm. Each structure is mutated by means of single point mutations, insertions, deletions and crossovers, to generate another population of structures for testing against the fitness function represented by a virtual receptor and possibly others such as ease of synthesis, toxicity etc as outlined above. The novelty of the approach is that the mutated structures together with a suitably defined fitness function and a genetic algorithm, can be used to explore very large areas of combinatorial space and generate lead structures likely to be active at the specified receptor. With suitable mutation strategies, virtually any part of the combinatorial universe (the region of compound space theoretically allowed by the rules of chemical bonding) can be explored. This means that areas devoid of useful biological activity would be rapidly vacated by the genetic algorithm, with more promising regions developing larger populations of lead structures. By reducing the level of similarity which will be tolerated, it may be possible to drive the population of chemical structures to high diversity whilst retaining useful lead activity.

6. Improved molecular multipole moment representation

Ultimately the quality of a virtual receptor is dependent on the quality of the molecular representation used to develop it. The quality of the virtual receptor is also dependent on the quality of the training data and possibly on the architecture of the neural net. In this regard, it is preferred that the numerical representation of the compound being analysed adequately represents the steric, electronic and lipophilic properties of the whole molecule.

A preferred representation, a schematic of which is illustrated in Figure 18, is the molecular multipole moment (MMM) representation, which is achieved by generating the zero-, first- and second-order molecular multi-pole moments (MMM) with respect to atomic mass and/or atomic charge. This characterisation is fundamentally a representation of the shape and/or charge of the compound as given by the moments of the shape, charge distributions

and/or molecular lipophilic moment (hydropoles). MMM descriptors relating solely to molecular shape are the three principal moments of inertia, Ix, Iy, Iz. The two descriptors that relate solely to charge are the magnitude of the dipole moment, p, and the magnitude of the principal quadrupole moment, Q. Descriptors that relate to shape and charge can be developed in a number of different ways. One example is by calculating the magnitudes of the dipolar components, the magnitudes of the components of displacement between the centre-of-mass and centre-of-dipole with respect to the principal inertia axes to provide the descriptors px, py, pz and dx, dy, dz. Quadrupolar components are calculated with respect to a translated inertial reference frame whose origin coincides with the centre-of-dipole, providing two additional descriptors Qxx and Qyy. This set of thirteen numbers is independent of the orientation and position of the molecules in three-dimensional space. (see B. D. Silverman and Daniel. E. Platt "Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR without Molecular Superposition" J. Med. Chem.,39 (11), 2129 -2140, 1996)

We disclose an improved method of generating molecular multipole moments as now described. In order to generate electrostatic, lipophilic or any other set of multipole moments, it is necessary to define a consistent set of axes. These axes should be invariant to a change in the elemental nature of an atom at any particular site, e.g. replacement of H by F. The axes should only vary their position or orientation when there is a structural change in the molecule, e.g. change in molecular conformation, an increase or decrease in the number of atoms making up the molecule, or some other structural change.

These requirements can be met using a set of principal pseudo moments of inertia. This is achieved by setting the mass of each atom in the molecule in this example to unity, or alternatively to a value reflecting the atom sizes, then evaluating the principal axis system in the traditional way used to determine principal moments of inertia.

The axes corresponding to these principal pseudo moments of inertia (steric multipoles in Figure 18) are then used as the basis for evaluating the vector and tensor components of the electrostatic, lipophilic or any other

property multipole moments.

The advantage of this method of axis definition is that molecules of similar shape will have their axis systems aligned when the molecules are put into atom-to-atom correspondence.

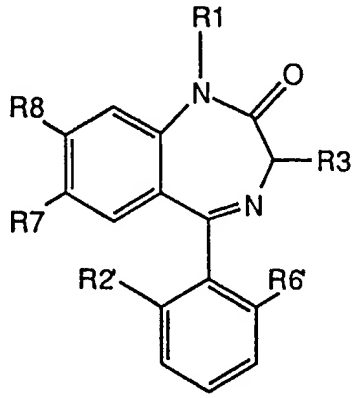
A lipophilic analogue of the steric and electrostatic multipole expansions may be derived by ascribing atomistic lipophilic values to each type of atom found in molecules. We did this by carrying out multiple regression analysis on a series of structures with known lipophilicities (described by the log of their water and octanol partition coefficients), using an atomistic representation (the numbers of each type of atom present in each molecule). This process may also be done in alternate ways as the literature suggests. We assumed that molecules could be attributed a lipophilic potential and proceeded to calculate lipophilic multipole moments (denoted as hydropoles) analogously to the way we determined the electrostatic multipoles.

7. Improved molecular eigenvalue index representation

The further aspect of the invention is an additional type of molecular representation. It is possible to describe the topographical relationships between atoms contained in a given molecular structure by means of connectivity or adjacency matrices. In general the diagonal elements of these matrices are zero and the off diagonal elements are unity only if the two atoms represented by the location of the matrix element are connected. Useful molecular representations may be derived from the eigenvalues of a modification of these matrices as first described by Burden (J. Chem. Inf. Comput. Sci., 29, 225 (1989). In this modification the connected, off diagonal elements are set equal to the square root of the bond order of the bonds connecting the two atoms represented by the matrix element, and the diagonal elements are set to empirically-derived values. Our invention involves the generation of useful molecular descriptors from eigenvalues of adjacency, or modified adjacency matrices in which the diagonal elements are values relating to steric, electrostatic or lipophilic properties of the constituents atoms of the compounds. Figure 19 illustrates a flow chart of a general process. In a

preferred embodiment it is envisaged that eigenvalues of three matrices (one each of steric, electrostatic, and lipophilic-related properties) are generated. The steric diagonal elements of the adjacency, or modified adjacency matrices could be the Van der Waals radii of the atoms; the electrostatic diagonal matrix elements could be the atom charges derived from empirical or molecular orbital calculations and; the lipophilic diagonal matrix elements could be the atomistic lipophilicities refereed to in the section above on molecular multipole moments.

Example 1



1,4-Benzodiazepin-2-ones

Experimental data illustrating the creation of a virtual receptor will now be provided, where the receptor is a benzodiazepine receptor (BZR) on a γ-aminobutyric acid receptor (GABAA) and is trained to recognise active compounds. Benzodiazepines have been used therapeutically to reduce anxiety, as tranquillisers, and also for their anticonvulsant effects in epilepsy. They act via the benzodiazepine receptor (BZR) on the GABAA.

The ANN used for this experiment had full connectivity, with the input layer of neurodes having linear transfer functions and all other layers of neurodes having sigmoidal transfer functions. In addition, the following neural network parameters were used:

- Momentum rate = 0.5
- Learning rate = 1.0
- Training patterns input noise = Gaussian (mean : 0; standard deviation :0.02)

The network calculations were performed using a commercial software package Propagator, however any neural network package could be used. The input data was scaled between 0 and 1, as it is between these values that the sigmoidal transfer functions range. Output data was also scaled appropriately.

The data set used was a set of 57 1,4-benzodiazepin-2-ones. This data set was chosen because their activity in relation to the receptor is known. The molecular representations of this data set that were employed are shown in Figure 1, while the size of training and validation sets and the number of networks generated during cross-validation is shown in Figure 2.

From figures 1 and 2, it is apparent that initial representations B1 and B2, which were based heavily on an atomistic approach, provided position information - for example, separate input parameters were provided for C4 atoms at positions 7, 1 and 3. In the case of B1, the representation comprised 25 input variables. In representation B2 the number of input parameters were slightly reduced by treating the halogens as being of the same element - "Hal". In the representations of B3 and B4, no positional information was provided - the neural network would not be told whether a C4 atom was attached to position 7, 1 or 3. The representation B4 differs from B3 in that it does not distinguish between the halogens.

The error of the total data set was recorded as the Standard Error of Prediction (SEP) defined as:

$$\text{SEP (Total data)} = \sqrt{\frac{\sum_{i}^{M} M_i (SEP)_i^2}{\sum_{i} M_i}}$$

The SEPval (which provides a measure of the predictive ability of the network) obtained from the two architectures used is shown in Figure 3A. A sample output from a 23:2:1 neural network using the B2 representation as input is shown in Figure 4A, and the sample output from an 11:4:1 neural network using the B3 representation is shown in Figure 4B.

Multiple Linear Regression (MLR) was performed on the data set to determine in a linear sense which input parameters were important in

modelling activity. MLR identified four linearly significant variables - C4, N3, O1 and Hal. These then formed the basis of the B5 representation. The results for each of these representations are shown in Figure 3B.

As a further variation of this example of the inventive concept, a functional group representation was used, which was independent of any reference to substituent position. The resultant representation is indicated in Figure 1 as A1. Due to this representation not being positionally dependent, the number of input parameters is much lower than the positionally dependent representations B1 and B2. Consequently, greater freedom is afforded in the architectures that can be devised. The results for the A1 representation are summarised in Figure 3C, whilst Figure 4C shows a typical output using the representation.

Multiple linear regression (MLR) was then performed on the data set, and four variables were identified as being linearly significant - C4, N, NO2, and Hal. These four variables then formed the basis of the A2 representation. The results obtained for this representation are summarised in Figure 3C.

Example 2

As a further variation of this example of the creation of a virtual receptor, a representation was used, namely MJ, which depicted the molecules based on their physicochemical properties. The representations used is explained in Figure 1, and the results obtained for this representation summarised in Figure 3D.

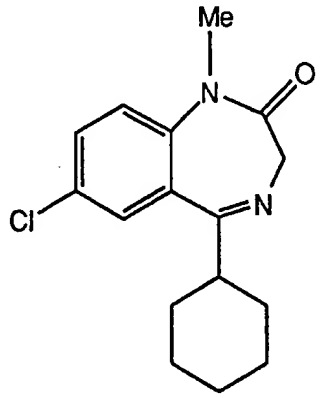
In Figure 5, the optimal architecture for each representation have been taken, and in each case R and R2 values calculated. The correlation coefficients provide a measure of how well the neural network has fit the entire data set - both training and validation sets. Thus, the R values provide a measure of the ability to predict the validation set, but also fit the training set.

From figure 5, it would appear that, in this experiment,the MJ representation fitted the data well, but the simpler representations had better predictive abilities. It also suggests that the atomistic and functional group representations are good at generalising outside of the training data.

Example 3

In a further example of creating a virtual receptor, a data set was compiled from the literature consisting of 321 compounds. These were broken up into two sets: 21 compounds would form the basis of training and validation sets. Training sets consisted of 270 compounds, validation sets consisted of 30 compounds. Thus, cross validation involved the generation of 10 training and validation set pairs.The neural network produced in each case was tested using the test set.

The representation used was based on the atomistic approach described previously. However, input parameters relating to the number and type of rings were added, thus affording the neural network some insight into the molecules topology. Twenty one input variables were used to represent each molecule: C(aromatic), C4, C3, C2, N(aromatic), N3, N2, N1, O2, O1, S, P, Cl, F, Br, I, 7-membered rings, 6-membered rings, 5-membered rings, 4-membered rings, 3-membered rings. An example of the representations is shown below:



Car	C4	C3	C2	Nar	N3	N2	N1	O2	O1	S	P
6	6	4	0	0	1	1	0	0	1	0	0

Cl	F	Br	I	7rings	6rings	5rings	4rings	3rings	IC50	pIC50
1	0	0	0	1	2	0	0	0	34	7.468

Unlike the previous example on benzodiazepines, where log IC50 was

modelled (IC50 being the binding affinity, which often corresponds to biological activity), this work modelled log 1/IC50 (known as the pIC50 value).

The results for this example are summarised in Figure 6 and a sample output from a 21:8:5:3:1 network is shown in Figure 7. From these results it is apparent that the 5 layer 21:8:5:3:1 network proved to be the most successful at modelling the data. It had the lowest SEPpred indicating that its ability to model the activity of compounds in both the validation and test sets was superior to any other architecture. In general the results were very good, with all architectures having an SEPpred lower than one pIC50 unit.

An indication of how successful the neural network has been in modelling the data set comes in comparing the above results with those obtained using MLR. MLR was performed on the data set twice - the first (MLR1) used only first-order terms, whilst the second (MLR2) used first and second order terms (but no cross-terms). MLR was employed on a "training set" of 270 compounds, then the resulting equation was tested on a validation set of 30 and a test set of 21. The results are compared with the neural network results on exactly the same data sets in Figure 8.

The results demonstrate that in both cases, the equation derived using MLR does not have as good predictive qualities as the neural networks.

Example 4

In order to evaluate the behaviour of the virtual receptor concept, a portion of a chemical structure database was screened and the biological activities of the members predicted. The database chosen was the first 7800 compounds in the Maybridge chemical database. While this database contains known, commercially available molecules, not hypothetical structures generated by techniques such as DBMaker, it serves to illustrate the screening procedure equally well.

The 7800 structures were converted into an atomistic representation similar to that outlined above. The representations were presented as input to a trained neural network representing a benzodiazepine receptor. Training was disabled so that the weights were fixed and the virtual receptor model generated 7800 outputs representing predicted log biological responses for

the 7800 compounds. The biological responses were ranked and the results are presented in Figure 9. The Maybridge column refers to the compound ID in the Maybridge database. The results of screening the benzodiazepine data set in the virtual receptor are also included.

Example 5

We carried out an analogous study to that in Example 1-3 to derive a Muscarinic Virtual Receptor from the analysis of a data set of 161 compounds which act upon the muscarinic receptor. Compounds capable of binding to this receptor (particularly the M1 muscarinic receptor in the cerebral cortex) are currently the subject of intense research, due to the believe that memory related problems in Alzheimer's disease could be treated using agonists at this receptor. The IC_{50} values sued in the analysis are the concentrations required to displace [^{3}H]Oxotremorine-M (OXO-M), an agonist at the M1 muscarinic receptor. The training sets contained 151 compounds, whilst the test set contained 10 compounds.

The analysis used 24 variables comprising 14 atomistic descriptors, 5 Randic indices, 5 Kier and Hall indices. PCA analysis reduced these to 17 principal components. The final neural network architecture used was 17 input nodes, 4 hidden nodes and 1 output node (i.e. 17:4:1 architecture). The range of dependent variable was 5.19 to 9.046. An example of observed versus calculated scaled log (activity) for training set (examples) is shown in figure 20.

The success of positionally-independent representations on the previous data set made it possible to represent the entire structure of a molecule (substituents and backbone) without the need to include positional information. This makes it possible to study structurally diverse data sets, and ascertain whether ANN's are capable of modelling activity in such data sets using simple representations.

The cross-validated Standard errors of prediction from regression analysis together with the training and test SEP from the Bayesian regularised neural net analysis are shown in Table 13. Unscaled, the SEP corresponded to an error of less than one pIC_{50} unit. The effective number of parameters from the Bayesian regularised artificial neural net training was 40 out of a possible 77

(number of neural net weights for 17:4:1 architecture). This gives a ratio of number of compounds to effective number of parameters of approximately 4 (this measure is referred to in the literature as the ρ value). The training r^2 value was 0.58 and the test r^2 value was 0.49.

Table 13. Cross-validated standard errors of prediction (SEPs) of Structurally Diverse Data Sets (scaled).

Method	Muscarinics
MLR	0.170
PCR	0.166
PLS1	0.166
BRANN (train)	0.137
BRANN (test)	0.134

Example 6

In order to evaluate the behaviour of the muscarinic virtual receptor concept, a portion of the National Cancer Institute (NCI) chemical structure database was screened and the biological activities of the members predicted. The database chosen was 110,000 compounds in the public domain version of the NCI chemical database. While this database contains known, commercially available molecules, not hypothetical structures generated by techniques such as DBMaker, it serves to illustrate the screening procedure equally well.

The 110,000 structures were converted into atomistic, Randic, and Kier and Hall representations similar to that outlined above. The representations were presented as input to a trained neural network representing a muscarinic receptor. Training was disabled so that the weights were fixed and the virtual receptor model generated outputs representing predicted log biological

responses for the compounds. The biological responses were ranked and a selected number of results are presented in Figure 12. The NCI column refers to the compound ID reported as the Chemical Abstracts Service (CAS) registry number. These best compounds from the database screening had -logIC50 (OXO-M) predicted to be approximately 1 nM.

NCI
50350-58-8
25288-35-1
5413-69-4
20135-35-7
24897-21-0
7401-38-9
6330-82-1
7147-80-0
3214-47-9
5398-51-6
17099-94-4
7356-48-1
5634-37-7
25349-34-2
31116-49-1
30885-73-5
25300-08-7

25300-09-8
25452-13-5
364-02-3
61719-87-7
19951-66-7

Example 7

This example describes the application of Bayesian Regularised Artificial Neural Nets to Virtual Receptors. We used diverse SAR data to develop neural net analogues of receptor SAR, resulting in models we believe can act as receptor surrogates or virtual receptors.

We used the 57 compound benzodiazepine data set employed in earlier examples to test the efficacy of BRANNs in QSAR and virtual receptor experiments. Four molecular indices were used; the well-studied Randic index; the valence modification to the Randic index by Kier and Hall(K); the atomistic (A); and an eigenvalue (E) index by Burden [Burden,F.R. Using Artificial Neural Networks to Predict Biological Activity from Simple Molecular Structural Considerations. Quant. Struct.-Activ.Relat., 15, 7-11 (1996);Burden, F.R. A Chemically Intuitive Molecular Index Based on the Eigenvalues of a Modified Adjacency Matrix. Quant.Struct.-Activ. Relat. 16, 309-14 (1997)].

These four types of index, R, K, A and E may be expected to complement each other. However, it is prudent to try and remove as much of the redundant information as possible before training neural networks. This is accomplished here by reducing the number of variables by principal component analysis (PCA). When principal components, which result from a linear transformation, are used prior to a non-linear regression, the usual criterion of ignoring those components with small variance is inappropriate. The number of principal

components that give the lowest standard error of prediction is the proper measure, and the use of a test set allows selection of models with best predictivity.

The ANN's used were three layer fully connected, feed forward networks which were trained using a Levenberg-Marquardt[Marquardt, D.W. J.Soc.Ind.Appl.Math. 11,431-441, (1963)] optimised back-propagation algorithm which incorporated Bayesian regularisation[MacKay,D.J.C. A Practical Bayesian Framework for Backprop Networks, Neural Computation,4,415-447,(1992)]. Using Bayesian regularisation removes the need to supply a validation set since it minimises a linear combination of squared errors and weights. It also modifies the linear combination so that at the end of training the resulting network has good generalisation qualities. It has also been suggested that there is no need for a test set since the application of the Bayesian statistics provides a network that has maximum generalisation. However, in this work we have provided a test set of 11 compounds for each calculation which has been selected by a K-means clustering algorithm.

The network architecture made use of 3 hidden nodes which proved to be more than sufficient in all cases with the Bayesian regularisation method estimating the number of effective parameters. The concerns about overfitting and overtraining are also removed by this method so that the production of a definitive and reproducible model is attained. The standard error of predictions (SEPs) and correlation coefficients, using the various representations, are shown in Table 14. A number of fully-connected ANN architectures, containing different numbers of hidden layers and nodes, were tested and a single hidden layer with 3 nodes was found to be optimal in each case. The number of effective parameters was always considerably less than the number of weights implied by the network architecture.

The data set compounds were scrambled to remove any inadvertent ordering effects such as by the magnitude of the biological activity. Before each network training, a K-means hierarchical clustering was carried out on the input variables and one compound from each cluster, at the 11 cluster level, was extracted for a test set. This test set, of 11 compounds, was not the same for

each set of input indices but was held to be consistent treatment of the data since the object of the study was to find a method for producing the best set of indices for future calculations.

Table 14 Bayesian Regularised Standard Errors of Prediction, SEP and Correlation Coefficients, R, of the various regressions: Data scaled 0 to 1.

Entry	Method	$N_I^{(a)}$	$N_{PC}^{(a)}$	SEP_{Train}	SEP_{Test}	R_{Test}	$N^{(b)}_{Par}$	$\rho eff^{(b)}$
1	$R^{(c)}$	5	4	0.155	0.190	0.639	7.30	6.03
2	$K^{(d)}$	5	4	0.182	0.187	0.858	4.60	9.56
3	$A^{(e)}$	11	8	0.098	0.121	0.866	19.21	2.29
4	$E^{(f)}$	10	9	0.123	0.121	0.768	23.8	1.85
5	RK	10	8	0.118	0.149	0.803	17.53	2.51
8	AE	21	10	0.100	0.129	0.847	18.92	2.32
9	RKA	21	15	0.062	0.092	0.928	26.4	1.67
10	RKEA	31	15	0.095	0.139	0.867	20.9	2.10

(a) NI = Number of independent variables.
NPC= Number of Principal Components used.

(b) NPar = Number of effective parameters.
peff= Number of input variables/NPar

(c) Randic [5] indices.

(d) Kier and Hall[6] indices.

(e) Atomistic Parameters
:H,C4,C3,C2,N3,N2,N1,O2,O1,F,Cl

(f) Eigenvalue indices.

The results show that the combination of two of the three descriptors can give a highly predictive QSAR model for the data; allowance being made for the experimental errors which are known to be small for this data set. In particular, the combination RKA may prove to be useful in screening very large data sets since the arithmetic operations in computing the descriptors are very fast. The eigenvalues depend on the diagonalisation of a matrix and are therefore rather slower as well as scaling as the cube of the number of atoms.

From these results it is apparent which molecules the virtual receptor model has predicted as being likely to exhibit the activity sought. It is to be noted that it is not intended that these results are precise results whereby the molecules definitely exhibit the activity sought: they are merely an indication of likely activity. This indication of likely activity combined with the speed within which the virtual receptor is able to scan large databases make the virtual receptor an extremely useful tool.

THE CLAIMS DEFINING THE INVENTION ARE AS FOLLOWS:

1. A method of screening a range of compounds, including:

(a) creating a virtual receptor;

(b) scanning the range of compounds with the virtual receptor so as to provide a measure of the likelihood of each compound exhibiting a particular characteristic.

2. A method as claimed in claim 1, wherein the range of compounds is in a real or virtual database.

3. The method of claim 1 or 2, wherein the virtual receptor is created to identify particular characteristics of compounds through the use of known compounds having the particular characteristics.

4. The method of claim 1, 2 or 3, wherein the virtual receptor is created by a conceptual model.

5. The method of any one of claims 1 to 4 wherein the range of compounds are represented in a molecular multipole moment format for input into the virtual receptor.

6. The method of any one of claims 1 to 4 wherein the range of compounds are represented in a molecular eigenvalue format for input into the virtual receptor.

7. A method of creating a virtual receptor capable of being used to scan a range of compounds and providing a measure indicative of whether the compounds are likely to exhibit a particular characteristic, including the steps of:

compiling a data set of compounds which exhibit the known characteristic;

forming an initial conceptual model with a given architecture;

converting the data set into a representation readable by the conceptual model;

training the conceptual model on at least a portion of the converted data set in order to improve the architecture of the conceptual model.

8. The method as claimed in claim 4 or 7, wherein the conceptual model is neural network technology.

9. The method of claim 7, wherein the representation of molecular structures in the data set is a molecular multipole moment representation.

10. The method of claim 7, wherein the representation of molecular structures in the data set is a molecular eigenvalue representation.

11. The method of claim 7, 8, 9 or 10, wherein the training includes across-validation step.

12. The method of claim 11, wherein the cross-validation process is eliminated or greatly reduced through use of a Bayesian regularised artificial neural net.

13. A method as claimed in claim 6, wherein the step of forming is based on a MML method, including Bayesian inference as exemplified as incorporated in a Bayesian neural net.

14. A method of calculating the measure of the likelihood of a compound exhibiting a particular characteristic, the method including the step of applying a 'fitness function'.

15. A method as claimed in claim 14, in which a virtual receptor is used as the fitness function.

16. A method as claimed in claim 14, where the fitness function is used in conjunction with a genetic evolutionary processes.

17. A method as claimed in claim 1 or 7, in combination with the method as claimed in claim 14.

18. A method as claimed in claim 17, further including the steps of:

selecting one or a number of compounds from the resultant population, and

applying a mutation operator in order to create children.

19. A method as claimed in claim 18, wherein the selection of one or a number of compounds is determined in accordance with at least one, or a combination of any of:

a valence function,
a stability function,
a safety function,
a biological function,
a molecular diversity function,
ease of synthesis function,
biological selectivity fitness,
cost fitness,
metabolic liability fitness,
combinatorial synthesis fitness,
pharmacokinetic efficiency fitness, and/or
chemical advantage efficiency.

20. A method as claimed in claim 18 or 19, further including the step of mutating a selected number of the population.

21. A method of deriving a virtual receptor by use of an artificial neural net.

22. A method of deriving a virtual receptor by use of a evolutionary neural network.

23. A virtual receptor rendered in a mathematical form.

24. A virtual receptor as claimed in claim 23, wherein an atomistic approach is used to generate the mathematical form, which classifies each atom according to its element and the number of connections.

25. A virtual receptor as claimed in claim 23, wherein a MMM approach is used to generate the mathematical form.

26. A virtual receptor as claimed in claim 23, wherein an Eigenvalue approach is used to generate the mathematical form.

27. A method of generating a virtual receptor by use of models which exhibit stability or compensate for noise.

28. A method as claimed in claim 27, in which a Bayesian regularised artificial neural network (BRANN) is the model used.

29. A method as claimed in claim 27, in which Maximum Entropy Method (MEM) is the model used.

30. An improved method of generating molecular multipole moments(MMM) adapted to generate electrostatic, lipophilic or any other set of multipole moments, wherein a consistent set of axes is defined, which are substantially invariant to a change in the elemental nature of an atom at any particular site.

31. A method of generating molecular descriptors from eigenvalues of adjacency, or modified adjacency matrices in which the diagonal elements are values relating to steric, electrostatic or lipophilic properties of the constituent

atoms of the compounds.

32. A method as claimed in claim 31, wherein the eigenvalues of three matrices (one each of steric, electrostatic, and lipophilic-related properties) are generated.

33. A method as claimed in claim 32, wherein the steric diagonal elements of the adjacency, or modified adjacency matrices are the Van der Waals radii of the atoms; the electrostatic diagonal matrix elements are the atom charges derived from empirical or molecular orbital calculations and; the lipophilic diagonal matrix elements are the atomistic lipophilicities refereed to in the section above on molecular multipole moments.

34. A method of creating a new or virtual chemical structure by use of mutation and/or cross-over operation.

35. A method as claimed in claim 34 wherein the mutation operation is applied to a SMILE string.

36. A method and/or apparatus as herein disclosed.

Representation		Input Parameters and Data Set
B1	•	Input Parameters : R7 : H, C4, C3, C2, N3, N2, N1, O2, O1, F, C1 R1 : H, C4, O2 R2': H, F, C1 R6': H, F, C1 R3 : H, C4, O2 R8 : H, C1
	•	No. of Compounds : 50. Compounds Removed : Ro 07-9957, Ro 05-3590, Halazepam, Ro 13-3780, Ro 06-7263, Ro 20-7078, Ro 20-8895
B2	•	Input Parameters : R7 : H, C4, C3, C2, N3, N2, N1, O2, O1, Hal R1 : H, C4, Hal R2': H, C4, Hal R6': H, Hal R3 : H, C4, O2 R8 : H, Hal
	•	No. of Compounds : 53. Compounds Removed : Halazepam, Ro 06-7263, Ro 20-7078, Ro 20-8895
B3	•	Input Parameters : H, C4, C3, C2, N3, N2, N1, O2, O1, F, C1
	•	No. of Compounds : 55. Compounds Removed : Ro 07-9957, Ro 13-3780
A1	•	Input Parameters : C4, C3, C2, N, NO_2, CO, OH, Hal, Degrees of Freedom
	•	No. of Compounds : 55. Compounds Removed : Ro 14-3074, Ro 06-9098
A2	•	Input Parameters : C4, N, NO_2, Hal
	•	No. of Compounds : As for A1,
B4	•	Input Parameters : H, C4, C3, C2, N3, N2, N1, O2, O1, Hal
	•	No. of Compounds : As for A1
B5	•	Input Parameters : C4, N3, O1, Hal
	•	No. of Compounds : As for A1
MJ	•	Input Parameters* : As for A1 π(R7), S(R7), MR(R1), $\Re$(R1), μ(R1), MR(R2'), $\Re$(R2'), MR(R6'), σ_m(R3), σ_p(R8)
	•	No. of Compounds : As for A1

* Where π = lipophilicity, MR = molar refractivity, S = polar constant, $\Re$ = resonance constant, μ = aromatic group dipole, σ_m = Hammett meta constant, σ_p = Hammett para constant

Fig1.

Representation	Size of Data Set	Training Set	Validation Set	No. of Networks
B1	50	45	5	10
B2	53	48(50)	5(3)	11
B3	55	50	5	11
A1	55	50	5	11
A2	55	50	5	11
B4	55	50	5	11
B5	55	50	5	11
MJ	55	50	5	11

Fig 2.

SEP_{val} values for B1, B2 and B3 representations.

Representation	Architecture	SEP_{val} [a]	ρ
B1	25 : 2 : 1	0.142 (0.393)	0.91
	25 : 2 : 1 : 1	0.129 (0.357)	0.88
B2	23 : 2 : 1	0.108 (0.299)	1.12
	23 : 3 : 1	0.116 (0.321)	0.75
B3	11 : 3 : 1	0.137 (0.379)	1.38
	11 : 4 : 1	0.117 (0.323)	1.04

Fig 3a.

[a] Scaled between 0.01 and 0.99. Unscaled values are in parentheses.

Results for functional group based representations.

Representation	Architecture	SEP_{val} [a]	ρ
A1	9 : 2 : 1	0.149 (0.412)	2.39
	9 : 3 : 1	0.125 (0.346)	1.62
	9 : 4 : 1	0.107 (0.296)	1.22
	9 : 3 : 2 : 1	0.112 (0.310)	1.34
	9 : 4 : 2 : 1	0.129 (0.357)	1.04
A2	4 : 4 : 1	0.118 (0.327)	2.20
	4 : 5 : 1	0.115 (0.318)	1.77
	4 : 3 : 2 : 1	0.114 (0.316)	2.12
	4 : 4 : 2 : 1	0.117 (0.324)	1.67
	4 : 5 : 2 : 1	0.114 (0.316)	1.38
	4 : 6 : 2 : 1	0.117 (0.324)	1.17

Fig 3b.

[a]Scaled between 0.01 and 0.99. Unscaled values are in parentheses.

Results for atomistic representation.

Representation	Architecture	SEP_{val} [a]	ρ
B4	9 : 2 : 1	0.139 (0.385)	2.39
	9 : 3 : 1	0.105 (0.291)	1.62
	9 : 4 : 1	0.106 (0.293)	1.22
	9 : 3 : 2 : 1	0.133 (0.368)	1.34
	9 : 4 : 2 : 1	0.129 (0.357)	1.04
B5	4 : 4 : 1	0.115 (0.318)	2.20
	4 : 5 : 1	0.113 (0.313)	1.77
	4 : 3 : 2 : 1	0.122 (0.338)	2.12
	4 : 4 : 2 : 1	0.124 (0.343)	1.67
	4 : 5 : 2 : 1	0.112 (0.310)	1.38
	4 : 5 : 3 : 1	0.126 (0.349)	1.17
	4 : 6 : 2 : 1	0.126 (0.349)	1.17

Fig 3c.

[a]Scaled between 0.01 and 0.09. Unscaled values are in parentheses.

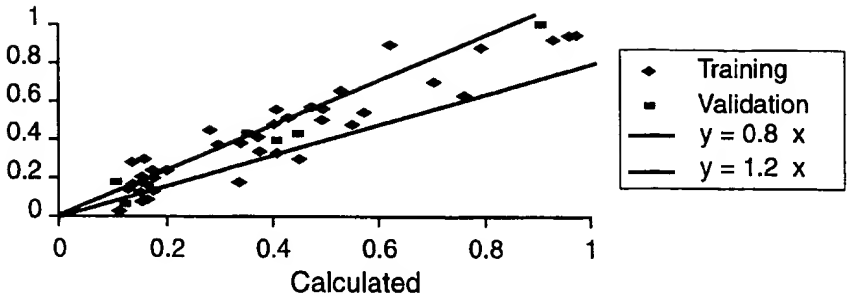
Results for physicochemical representation.

Representation	Architecture	$RMSE_{val}$ [a]	ρ
MJ	10 : 2 : 1	0.129 (0.357)	2.20
	10 : 3 : 1	0.115 (0.318)	1.49
	10 : 4 : 1	0.115 (0.318)	1.12
	10 : 3 : 2 : 1	0.117 (0.324)	1.25
	10 : 4 : 2 : 1	0.123 (0.341)	0.96

Fig 3d.

[a]Scaled between 0.01 and 0.09. Unscaled values are shown in parentheses

Fig 4a.


1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
Calculated
Training
Validation
y = 0.8 x
y = 1.2 x

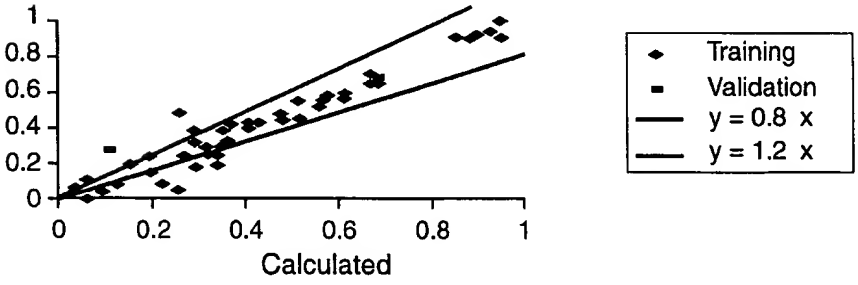

1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
Calculated
Training
Validation
y = 0.8 x
y = 1.2 x

Fig 4c.

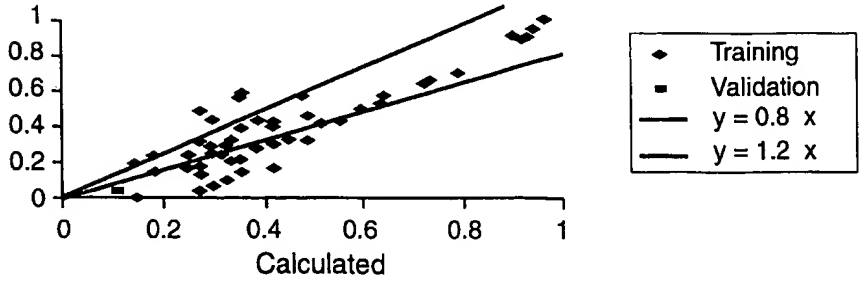

1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
Calculated
Training
Validation
y = 0.8 x
y = 1.2 x

Fig 5.

Representation	$RMSE_{val}$	R	R^2
A1 (9 : 4 : 1)	0.107 (0.296)	0.900149	0.811514
A2 (4 : 5 : 2 : 1)	0.114 (0.316)	0.876040	0.769032
B4 (9 : 3 : 1)	0.105 (0.291)	0.885573	0.787354
B5 (4 : 5 : 2 : 1)	0.112 (0.310)	0.837073	0.703643
MJ (10 : 4 : 1)	0.115 (0.318)	0.917694	0.844229

Fig 6.

Summary of scaled SEP results on structuraly diverse benzodiazepine data set. Unscaled SEP values (in pIC_{50} units) are shown in parentheses.

Architecture	SEP(Train)	SEP(Valid[n])	SEP(Test)	SEP (Pred)[a]	SEP(Ave)[b]
21:3:3:2:1	0.128	0.125	0.163	0.142	0.131
	(0.691)	(0.674)	(0.880)	(0.766)	(0.704)
21:3:2:1	0.164	0.128	0.1662	0.145	0.161
	(0.885)	(0.691)	(0.896)	(0.782)	(0.870)
21:3:1	0.132	0.126	0.161	0.142	0.133
	(0.712)	(0.682)	(0.870)	(0.765)	(0.721)
21:6:3:2:1	0.116	0.121	0.165	0.141	0.120
	(0.626)	(0.654)	(0.892)	(0.761)	(0.649)
21:6:4:1	0.110	0.127	0.171	0.147	0.117
	(0.593)	(0.685)	(0.920)	(0.790)	(0.629)
21:7:1	0.105	0.120	0.172	0.144	0.112
	(0.565)	(0.647)	(0.930)	(0.776)	(0.603)
21:8:5:3:1	0.101	0.118	0.167	0.141	0.108
	(0.546)	(0.637)	(0.903)	(0.758)	(0.585)
21:8:5:3:1[c]	0.101	0.118	0.115	0.117	0.104
	(0.546)	(0.637)	(0.618)	(0.629)	(0.560)
21:9:4:1	0.110	0.120	0.172	0.144	0.116
	(0.595)	(0.647)	(0.929)	(0.776)	(0.627)
21:11:1	0.109	0.125	0.178	0.149	0.116
	(0.587)	(0.675)	(0.958)	(0.804)	(0.627)

[a] The average SEP value of the validation and test sets.

[b] The average SEP value of the training, validation and test sets.

[c] Modified test set - the largest outlier was removed from the test set.

Fig 7.

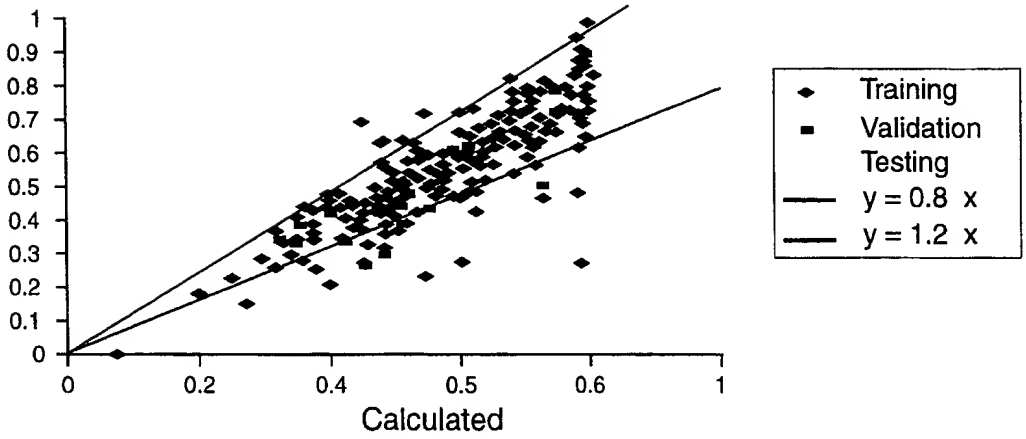


Fig 8.

Method	RMSE (VAL)	RMSE (Test)
ANN	0.137	0.147
21:6:3:2:1	(0.739)	(0.793)
ANN	0.101	0.119
21:8:5:3:1	(0.545)	(0.642)
MRL 1	0.158	0.150
	(0.852)	(0.809)
MRL 2	0.324	0.211
	(1.747)	(1.138)

Fig 9.

Log IC_{50} values for virtual benzodiazepine receptor.

Maybridge	VR pI_{50}	M.W.	Bz Dataset	VR pI_{50}	M.W.
3553	9.624	436	165	9.307	477
3547	9.606	436	15	9.000	393
2967	9.600	228	185	8.973	407
5404	9.585	493	136	8.896	386
4604	9.579	432	125	8.890	393
3549	9.565	452	217	8.875	403
1400	9.563	476	197	8.849	427
2839	9.562	457	121	8.819	369
3550	9.558	444	127	8.794	364
4065	9.553	470	196	8.792	413
3232	9.533	468	126	8.741	440
3029	9.526	470	195	8.712	399
5737	9.521	466	213	8.712	399
2307	9.519	494	214	8.712	399
5738	9.519	494	134	8.703	371
6726	9.512	493	115	8.700	342
3548	9.503	410	164	8.696	388
4719	9.501	497	183	8.693	417
3556	9.492	453	124	8.692	349
3557	9.492	453	163	8.676	328
3540	9.485	446	145	8.668	341
3563	9.483	382	218	8.651	450
5405	9.480	459	177	8.647	385
3555	9.479	425	317	8.618	300
3634	9.475	442	184	8.609	464
5590	9.461	447	225	8.605	314
2980	9.461	451	319	8.605	314
5740	9.458	475	146	8.601	353.5
7690	9.455	480.2	12	8.585	339.4
2295	9.448	432.1	181	8.576	391.5

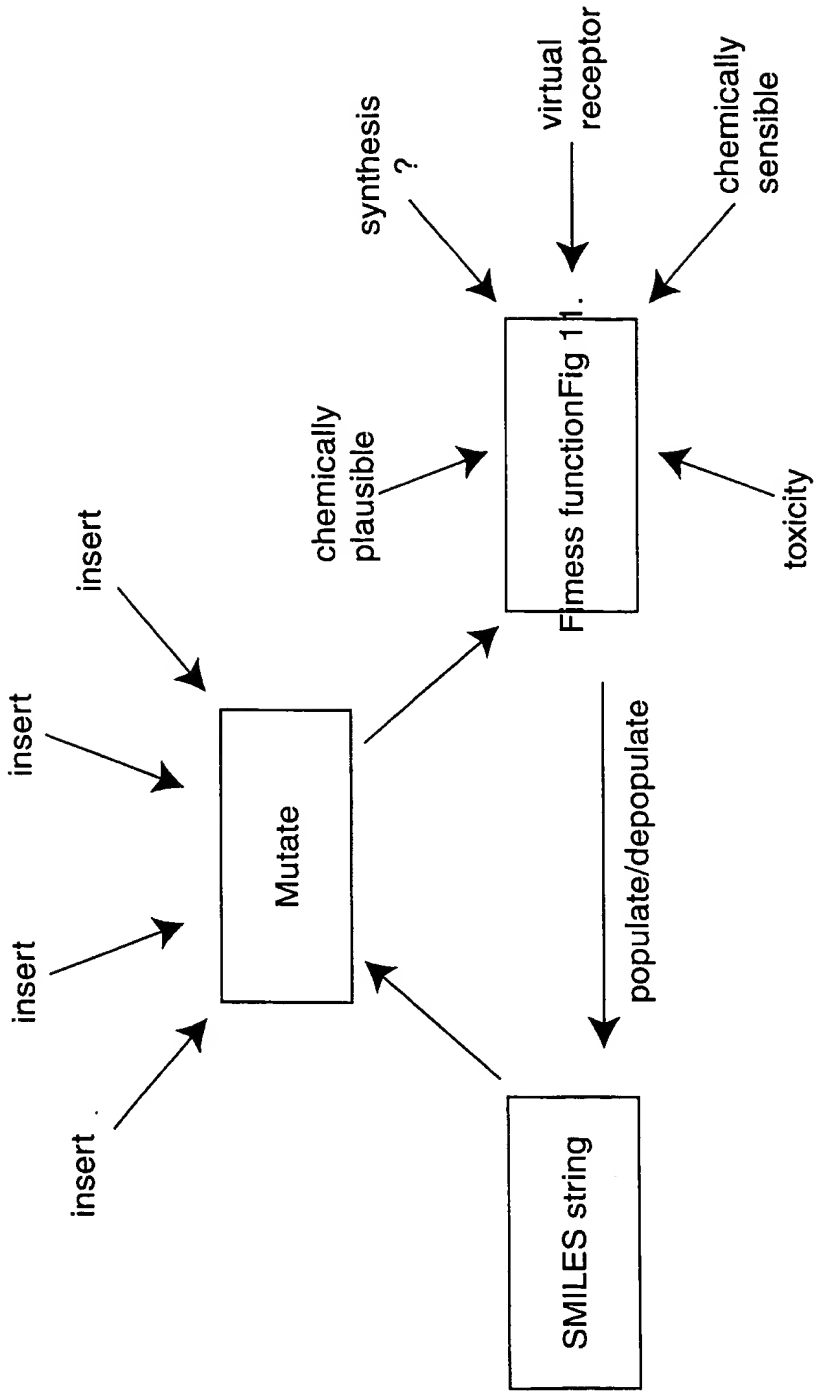

synthesis ?
virtual receptor
chemically sensible
chemically plausible
Fimess functionFig 1
toxicity
insert
insert
insert
insert
Mutate
populate/depopulate
SMILES string

Fig 10.

Fig 11.

the Cycle of Reproduction

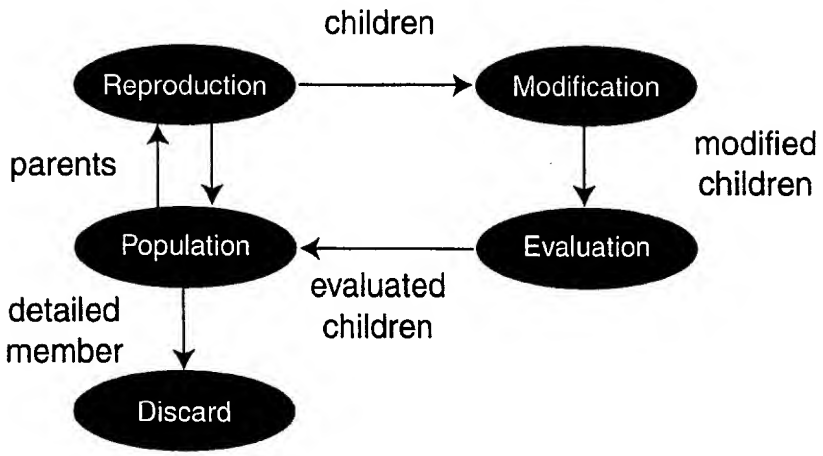


Fig 12.

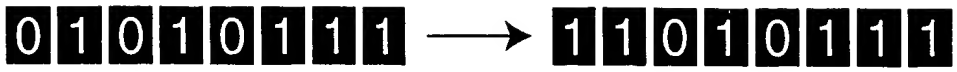


Fig 13.

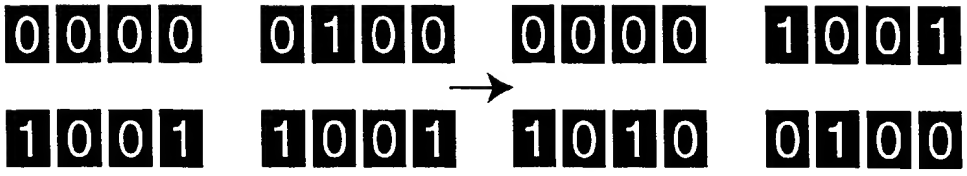


Fig 14.

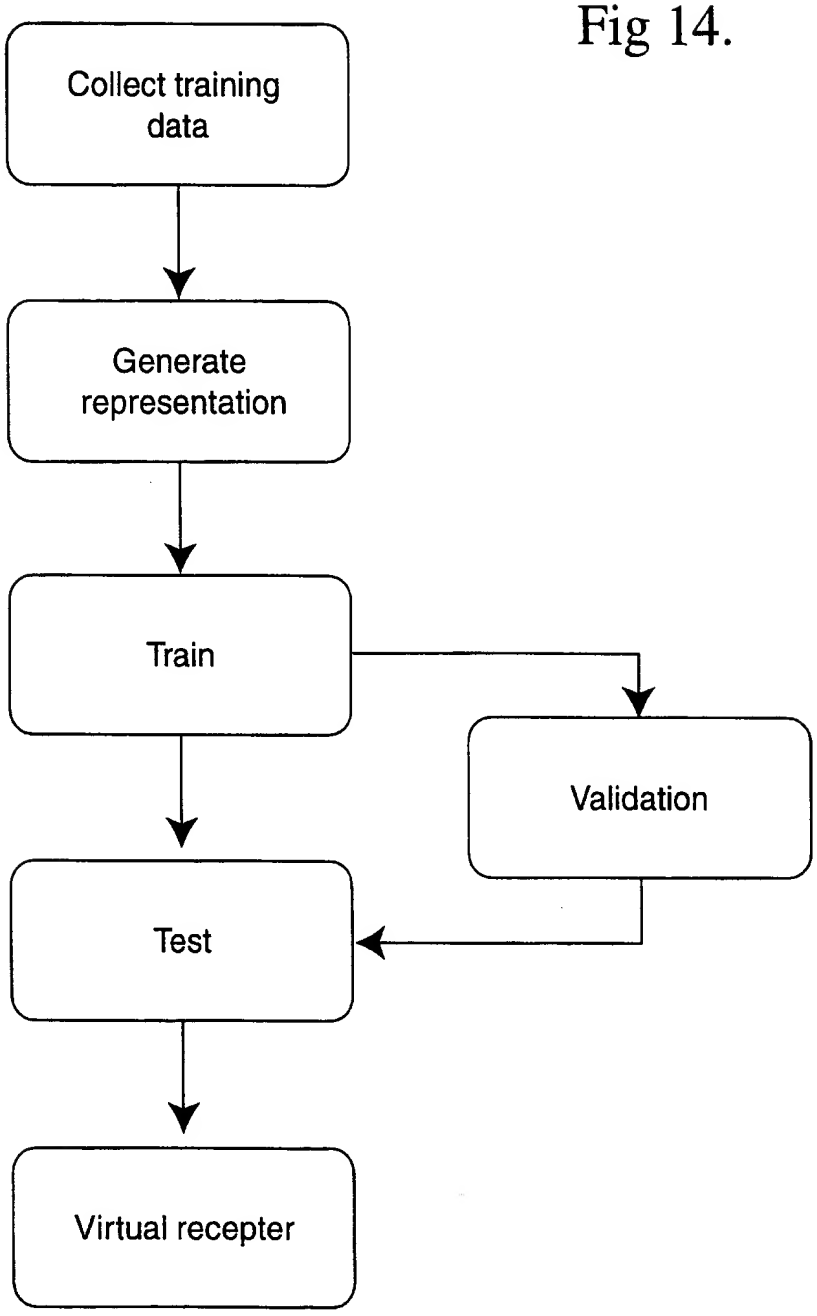

Collect training data
Generate representation
Train
Validation
Test
Virtual recepter

Fig 15.

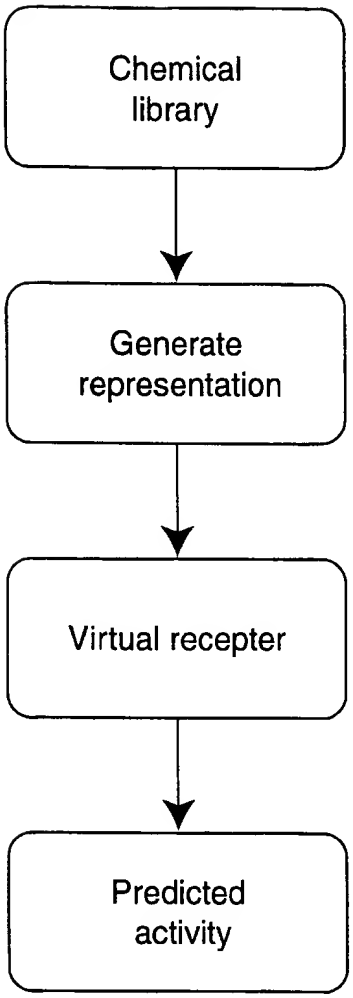

Chemical library
Generate representation
Virtual recepter
Predicted activity

Fig 16.

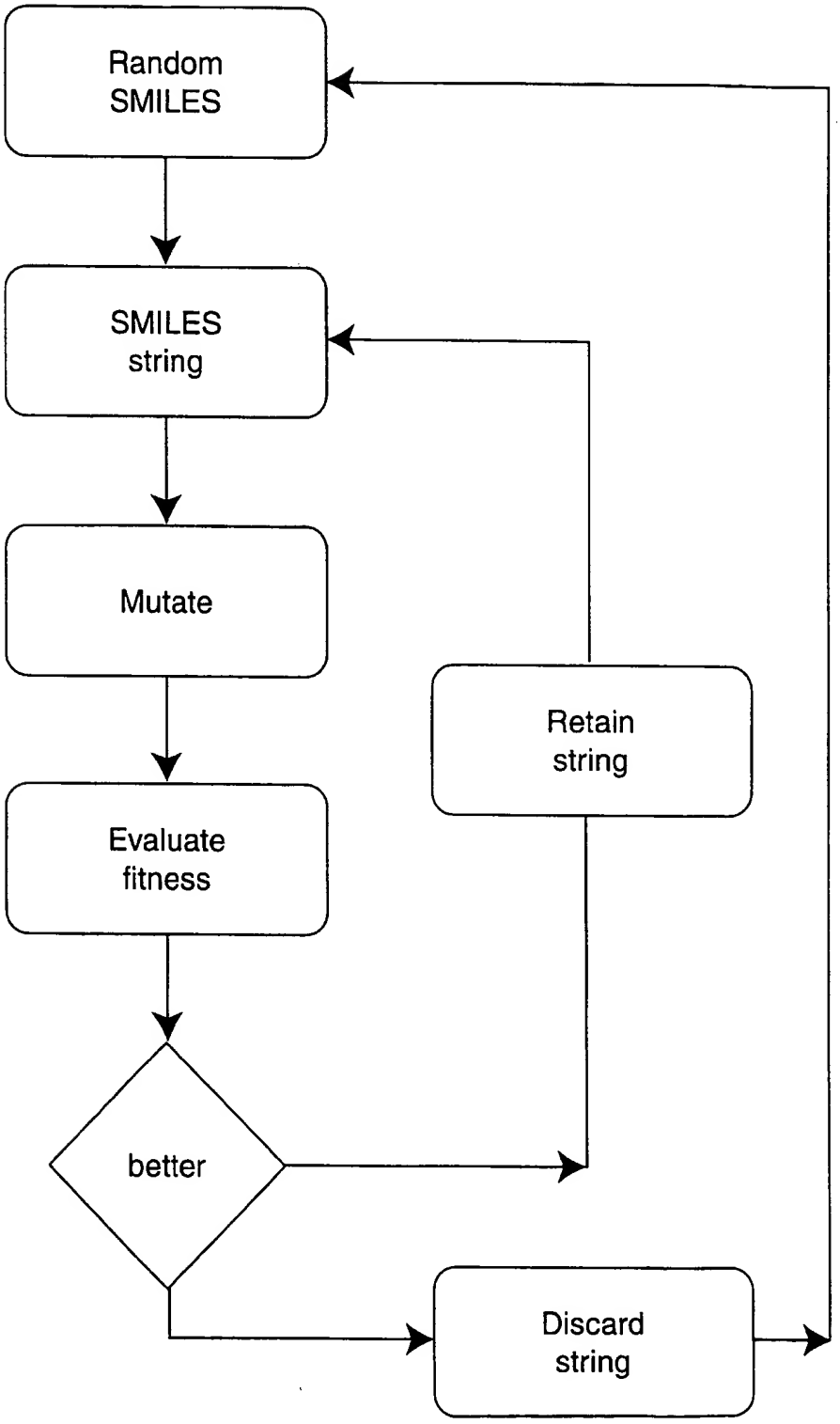

Random SMILES
SMILES string
Mutate
Evaluate fitness
better
Retain string
Discard string

Fig 17.

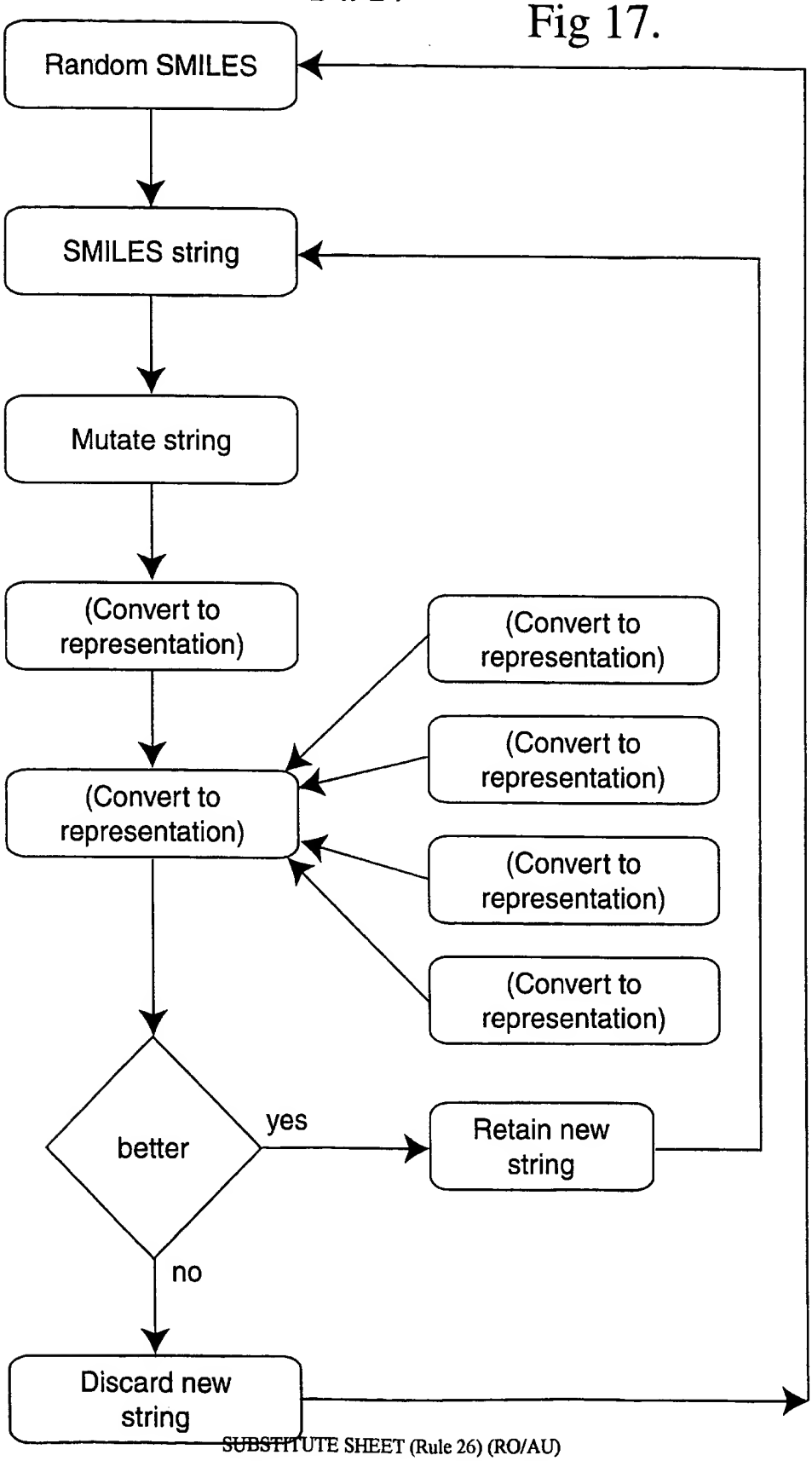

Random SMILES
SMILES string
Mutate string
(Convert to representation)
(Convert to representation)
(Convert to representation)
(Convert to representation)
(Convert to representation)
(Convert to representation)
better
yes
Retain new string
no
Discard new string

Fig 18.

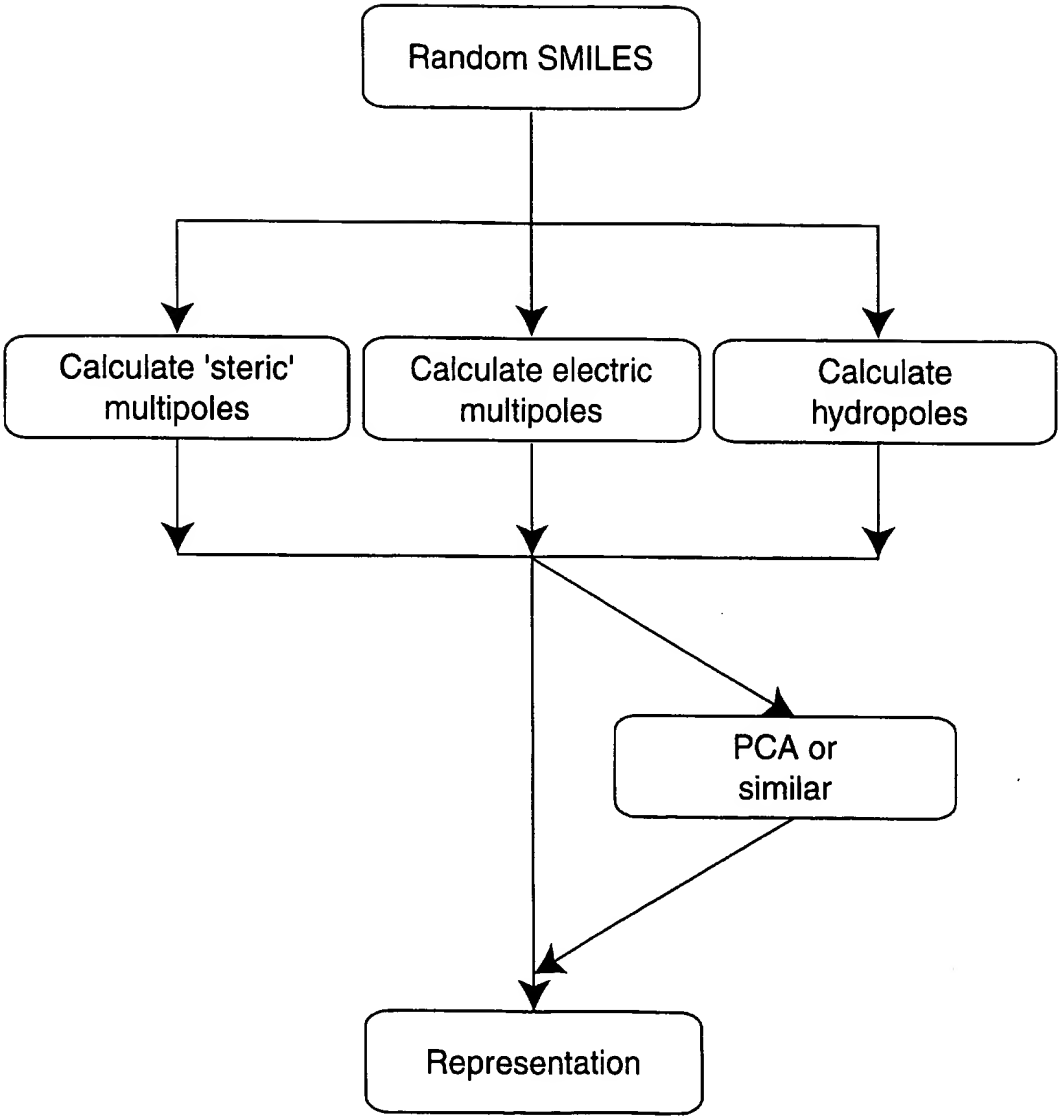
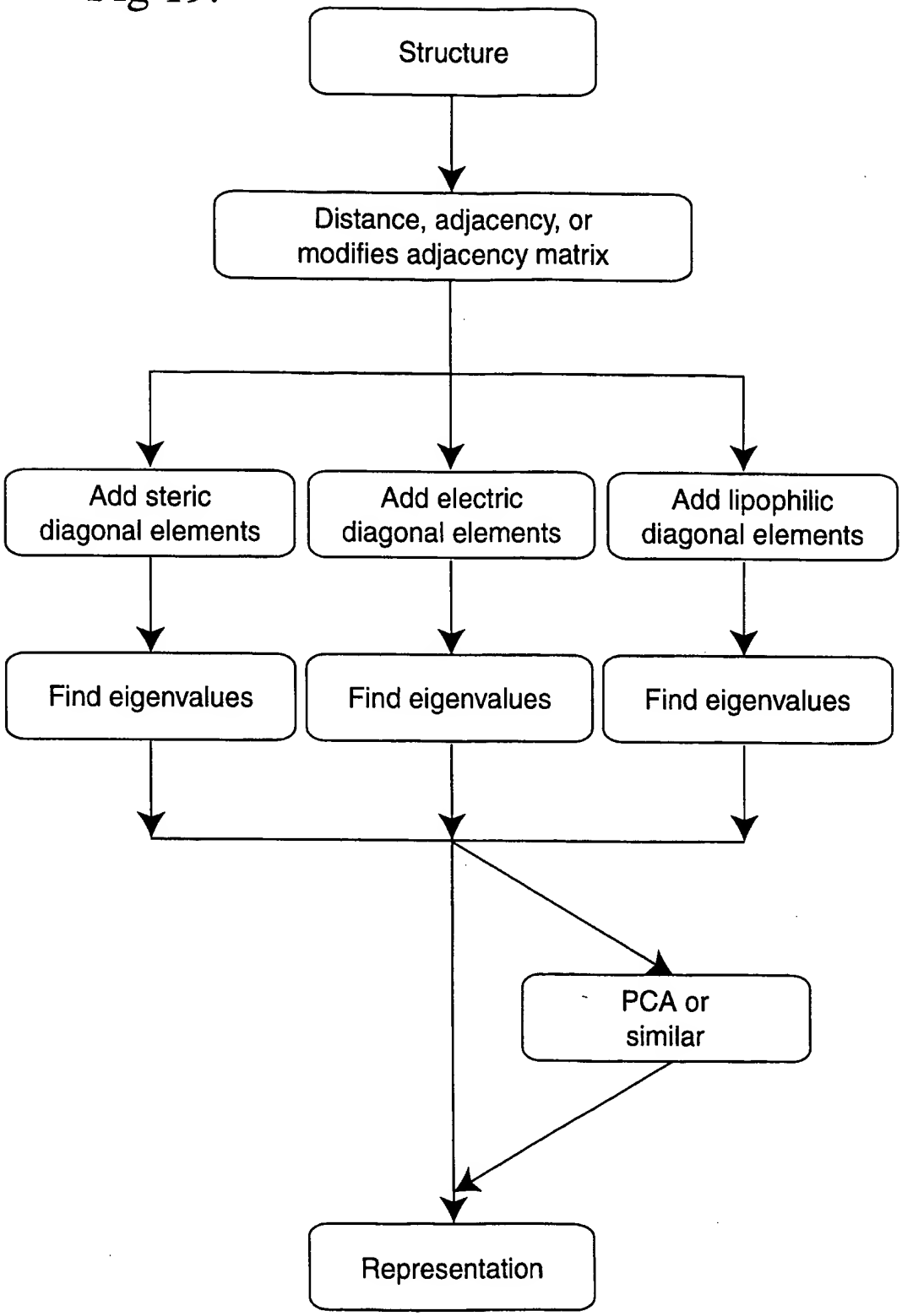
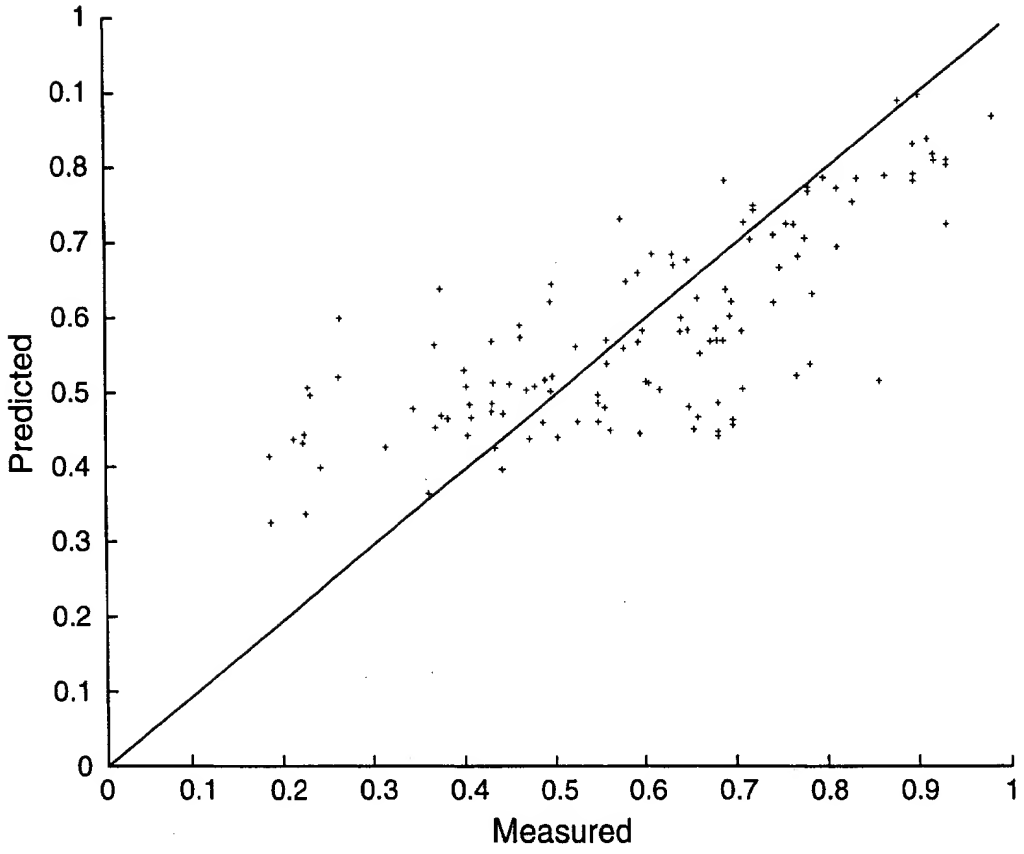

Random SMILES
Calculate 'steric' multipoles
Calculate electric multipoles
Calculate hydropoles
PCA or similar
Representation

Fig 19.


Structure
Distance, adjacency, or modifies adjacency matrix
Add steric diagonal elements
Add electric diagonal elements
Add lipophilic diagonal elements
Find eigenvalues
Find eigenvalues
Find eigenvalues
PCA or similar
Representation

17/17

Fig 20.



INTERNATIONAL SEARCH REPORT

International Application No.
PCT/AU 98/00715

A. CLASSIFICATION OF SUBJECT MATTER

Int Cl[6]: G06F 159:00, G06F 15/18

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int Cl.[5] - G06F 15/42, Int Cl.[6] G06F 15/18, G06F 159:00, C12Q, C12M

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
WPAT:- RECEPTOR#, ACTIVE ( ) SITE#, NEURAL ( ) NET:

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	US 5699268 (SCHMIDT) 16 December 1997	
P,X	see whole document	1-4, 23, 24, 27
X	see whole document	14-20, 34
Y	see whole document	7, 8, 21, 22
	US 5526281 (CHAPMAN ET AL) 11 June 1996	
Y	see whole document	7, 8, 21, 22

[X] Further documents are listed in the continuation of Box C

[X] See patent family annex

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
5 November 1998	10 NOvember 1998
Name and mailing address of the ISA/AU AUSTRALIAN PATENT OFFICE PO BOX 200 WODEN ACT 2606 AUSTRALIA Facsimile No.: (02) 6285 3929	Authorized officer **J.W. THOMSON** Telephone No.: (02) 6283 2214

INTERNATIONAL SEARCH REPORT

International Application No.
PCT/AU 98/00715

C (Continuation) **DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5459077 (MOORE ET AL) 17 October 1995 see whole document	1-5, 14, 15, 17, 23-25, 27, 34
Y	see whole document	7-9, 21, 22, 30
Y	US 5524086 (KIYUNA ET AL) 4 June 1996 see whole document	7-9, 21, 22, 30
X	"Molecular Biology and Molecular Simulations" in Solutions Magazine, pub April 1997 at http: // www.biosym.com/about/solutions mag/index.html see whole document	1-4, 14, 15, 17, 23, 24, 27, 34

INTERNATIONAL SEARCH REPORT

International Application No.
PCT/AU 98/00715

Box 1 Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This International Search Report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a)

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. Claims 1-33, 35-36
A virtual representation of a chemical receptor site

2. Claim 34
Creating new chemical structures by mutation.

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims

2. ☒ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International Application No.
PCT/AU 98/00715

This Annex lists the known "A" publication level patent family members relating to the patent documents cited in the above-mentioned international search report. The Australian Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

Patent Document Cited in Search Report		Patent Family Member					
US	5526281	AU	73119/94	WO	94/28504	US	5703792
US	5459077	AU	70347/91	CA	2072363	EP	557276
		HU	9202147	HU	65361	JP	5503691
		PT	96440	WO	91/10140	ZA	9010460
US	5524086	CA	2123563	JP	6319713	JP	2739804

END OF ANNEX